PB86-139516

Consolidation of Baseline Information Development of Methodology, and Investigation of Thermal Impacts on Freshwater Shellfish, Insects, and Other Biota

(U.S.) Tennessee Valley Authority, Muscle Shoals
Muscle Shoals, AL

Prepared for

Environmental Protection Agency, Washington, DC

Mar 84

U.S. Department of Commerce
National Technical Information Service
NTIS

TVA

Tennessee
Valley
Authority

Office of Natural
Resources
Muscle Shoals AL 35660

PB86-139516

EPA

United States
Environmental Protection
Agency

Office of Environmental
Processes and Effects Research
Washington DC 20460

EPA-600/7-84-042
March 1984

Research and Development

Consolidation of Baseline Information, Development of Methodology, and Investigation of Thermal Impacts on Freshwater Shellfish, Insects, and Other Biota

Interagency Energy/Environment R&D Program Report

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO.
EPA/600/7-84/042

2.

3. RECIPIENT'S ACCESSION NO.
PB86 T39516/AS

4. TITLE AND SUBTITLE
CONSOLIDATION OF BASELINE INFORMATION, DEVELOPMENT OF METHODOLOGY, AND INVESTIGATION OF THERMAL IMPACTS ON FRESHWATER SHELLFISH, INSECTS, AND OTHER BIOTA

5. REPORT DATE
March 1984

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
John S. Crossman, James B. Wright, Jr., and
Roger L. Kaesler

8. PERFORMING ORGANIZATION REPORT NO.
TVA/EP-78/09

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Office of Natural Resources
Tennessee Valley Authority
Knoxville, TN 37902

10. PROGRAM ELEMENT NO.
INE-625A

11. CONTRACT/GRANT NO.
EPA-IAG-DS-E721

12. SPONSORING AGENCY NAME AND ADDRESS
Offices of Environmental Processes and Effects Research
Office of Research and Development
U.S. Environmental Protection Agency
Washington, DC 20460

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE
EPA/600/16

15. SUPPLEMENTARY NOTES

16. ABSTRACT

A computerized information system was developed for storing, retrieving, and analyzing data collected during limnological surveys. To facilitate storage of information, a series of hierarchial codes was developed. These codes not only reduced storage requirements, but also helped reduce computing costs.

The information system utilized three analytical procedures, cluster analysis, ordination using nonmetric multidimensional scaling (MDS), and measurement of species diversity from information theory.

Results indicated that identification to species contributed little information about the structure of communities that discrimination of genera had not already provided.

The heuristic properties of species diversity were used to evaluate two classifications stressing fucntional morphology and trophic-functional relationships of benthic invertebrates, independent of the taxonomic hierarchy. Both methods produced results similar to ones obtained by cluster analysis, suggesting that they merit further investigation.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT
RELEASE TO PUBLIC

19. SECURITY CLASS *(This Report)*
UNCLASSIFIED

20. SECURITY CLASS *(This page)*
UNCLASSIFIED

21. NO. OF PAGES

22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

INSTRUCTIONS

1. **REPORT NUMBER**
Insert the EPA report number as it appears on the cover of the publication.

2. **LEAVE BLANK**

3. **RECIPIENTS ACCESSION NUMBER**
Reserved for use by each report recipient.

4. **TITLE AND SUBTITLE**
Title should indicate clearly and briefly the subject coverage of the report, and be displayed prominently. Set subtitle, if used, in smaller type or otherwise subordinate it to main title. When a report is prepared in more than one volume, repeat the primary title, add volume number and include subtitle for the specific title.

5. **REPORT DATE**
Each report shall carry a date indicating at least month and year. Indicate the basis on which it was selected *(e.g., date of issue, date of approval, date of preparation, etc.)*.

6. **PERFORMING ORGANIZATION CODE**
Leave blank.

7. **AUTHOR(S)**
Give name(s) in conventional order *(John R. Doe, J. Robert Doe, etc.)*. List author's affiliation if it differs from the performing organization.

8. **PERFORMING ORGANIZATION REPORT NUMBER**
Insert if performing organization wishes to assign this number.

9. **PERFORMING ORGANIZATION NAME AND ADDRESS**
Give name, street, city, state, and ZIP code. List no more than two levels of an organizational hirearchy.

10. **PROGRAM ELEMENT NUMBER**
Use the program element number under which the report was prepared. Subordinate numbers may be included in parentheses.

11. **CONTRACT/GRANT NUMBER**
Insert contract or grant number under which report was prepared.

12. **SPONSORING AGENCY NAME AND ADDRESS**
Include ZIP code.

13. **TYPE OF REPORT AND PERIOD COVERED**
Indicate interim final, etc., and if applicable, dates covered.

14. **SPONSORING AGENCY CODE**
Insert appropriate code.

15. **SUPPLEMENTARY NOTES**
Enter information not included elsewhere but useful, such as: Prepared in cooperation with, Translation of, Presented at conference of, To be published in, Supersedes, Supplements, etc.

16. **ABSTRACT**
Include a brief *(200 words or less)* factual summary of the most significant information contained in the report. If the report contains a significant bibliography or literature survey, mention it here.

17. **KEY WORDS AND DOCUMENT ANALYSIS**
(a) DESCRIPTORS - Select from the Thesaurus of Engineering and Scientific Terms the proper authorized terms that identify the major concept of the research and are sufficiently specific and precise to be used as index entries for cataloging.

(b) IDENTIFIERS AND OPEN-ENDED TERMS - Use identifiers for project names, code names, equipment designators, etc. Use open-ended terms written in descriptor form for those subjects for which no descriptor exists.

(c) COSATI FIELD GROUP - Field and group assignments are to be taken from the 1965 COSATI Subject Category List. Since the majority of documents are multidisciplinary in nature, the Primary Field/Group assignment(s) will be specific discipline, area of human endeavor, or type of physical object. The application(s) will be cross-referenced with secondary Field/Group assignments that will follow the primary posting(s).

18. **DISTRIBUTION STATEMENT**
Denote releasability to the public or limitation for reasons other than security for example "Release Unlimited." Cite any availability to the public, with address and price.

19. & 20. **SECURITY CLASSIFICATION**
DO NOT submit classified reports to the National Technical Information service.

21. **NUMBER OF PAGES**
Insert the total number of pages, including this one and unnumbered pages, but exclude distribution list, if any.

22. **PRICE**
Insert the price set by the National Technical Information Service or the Government Printing Office, if known.

EPA/600/7-84/042
March 1984

CONSOLIDATION OF BASELINE INFORMATION, DEVELOPMENT OF METHODOLOGY, AND INVESTIGATION OF THERMAL IMPACTS ON FRESHWATER SHELLFISH, INSECTS, AND OTHER BIOTA

by

John S. Crossman and James R. Wright, Jr.
Office of Natural Resources
Tennessee Valley Authority
Knoxville, Tennessee 37902

Roger L. Kaesler
Department of Geology
University of Kansas
Lawrence, Kansas 66045

Interagency Agreement EPA-IAG-DS-E721
Project No. E-AP-80-BDR
Program Element No. INE-625A

Project Officer

Alfred Galli
Office of Environmental Processes and Effects Research
U.S. Environmental Protection Agency
Washington, DC 20460

Prepared for

Office of Environmental Processes and Effects Research
Office of Research and Development
U.S. Environmental Protection Agency
Washington, DC 20460

DISCLAIMER

This report was prepared by the Tennessee Valley Authority and has been reviewed by the Office of Research and Development, Energy and Air Division, U.S. Environmental Protection Agency, and approved for publication. Although the research described in this document has been funded wholly or in part by the United States Environmental Protection Agency through Interagency Agreement No. EPA-IAG-82-D-X0511 with TVA, it has not been subject to Agency policy and peer review and therefore does not necessarily reflect the views of the agency or the Tennessee Valley Authority and no official endorsement should be inferred.

ABSTRACT

A computerized information system was developed for storing, retrieving, and analyzing data collected during limnological surveys. The system accommodates 19 variables and uses the Statistical Analysis System as the basic data-management system. To facilitate storage of information, a series of hierarchical codes was developed. These codes not only reduced storage requirements but also helped reduce computing costs.

When the information system had been developed, three analytical procedures were tested used in various forms, and evaluated as tools for analyzing benthic macroinvertebrate data sets. The first two of these were cluster analysis and ordination using nonmetric multidimensional scaling (MDS). Tests of these two methods included, first, development of a rationale for selecting methods of transforming the data and choosing coefficients of similarity to express relationships between samples and, second, application of the two methods to data from the Clinch River, Virginia, and the Cumberland River, Tennessee. The resulting dendrograms from cluster analysis and ordinations from multidimensional scaling were evaluated by preparing analytical tables and dealing with small subsets of the total data set.

Measurements of species diversity from information theory were the third analytical technique considered, again using benthic macroinvertebrate data. First, the relative importance of diversity at the species level was compared to components of diversity contributed by other categories in the taxonomic hierarchy. Results indicated that identifications to species contributed little information about the structure of the communities that discrimination of genera had not already contributed. Second, the heuristic properties of species diversity were used to evaluate two classifications stressing functional morphology and trophic-functional relationships of benthic invertebrates, independent of the taxonomic hierarchy. Both methods produced results similar to ones obtained by cluster analysis, suggesting that they merit further investigation.

CONTENTS

Page

Abstract iii
List of Figures vii
List of Tables xii
List of Abbreviations and Symbols xvi
Disclaimer ii
Acknowledgments xvii

1. Introduction 1

2. Conclusions and Recommendations 3
  2.1 Introduction 3
  2.2 Generic 3
  2.3 Cluster Analysis 3
  2.4 Ordination 4
  2.5 Indices of Diversity 5
  2.6 Final Statement 5

3. Development of the Information System 6
  3.1 General Description of the Information System 6
  3.2 Preparation of Data 6
  3.3 Editing of Data 12

4. Methods 31
  4.1 Description of Analytical Techniques 31
    4.1.1 Cluster Analysis 31
    4.1.2 Ordination by Nonmetric Multidimensional Scaling . . . 34
    4.1.3 Indices of Species Diversity 41
  4.2 Description of Data Sets 42
    4.2.1 Clinch River Data Set 42
    4.2.2 Cumberland River Data Sets--1973 and 1975 44

5. Cluster Analysis 54
  5.1 General Description 54
  5.2 Analytical Procedures 54
    5.2.1 Selection of Similarity Coefficients 54
    5.2.2 Reducing Size of Data Matrices 57
    5.2.3 Evaluation of Distortion 60
  5.3 Results 60
    5.3.1 Q-Mode Analysis 60
      5.3.1.1 Clinch River Data Set 60
        5.3.1.1.1 Presence-absence data 60
        5.3.1.1.2 Quantitative data, counts of species 62

CONTENTS (continued)

Page

5.3.1.2 Cumberland River Data Set--1973 62
5.3.1.2.1 Substrate 62
5.3.1.2.2 Zoomacrobenthos 62
5.3.1.2.2.1 Presence-absence data 65
5.3.1.2.2.2 Quantitative data, counts of species . 65
5.3.1.2.3 Summary 65

5. Cluster Analysis (continued)

5.3.1.3 Cumberland River Data Set--1975 69
5.3.1.3.1 Substratum 69
5.3.1.3.2 Zoomacrobenthos 69
5.3.1.3.2.1 Presence-absence data 69
5.3.1.3.2.2 Quantitative data, species counts . . 70
5.3.1.3.4 Summary 70
5.3.2 R-Mode Analysis 73
5.3.2.1 Introduction 73
5.3.2.2 Clinch River Data Set 76
5.3.2.2.1 Presence-absence data 76
5.3.2.2.2 Quantitative data, species counts 85
5.3.2.2.3 Summary 85
5.3.2.3 Cumberland River Data Set--1973 85
5.3.2.3.1 Presence-absence data 85
5.3.2.3.2 Quantitative data, species counts 87

6. Ordination--Nonmetric Multidimensional Scaling 124
6.1 General Description 124
6.2 Analytical Procedures 124
6.3 Results 125
6.3.1 Q-Mode Analysis--Clinch River Data 125
6.3.1.1 Presence-Absence Data 125
6.3.1.2 Quantitative Data (Species Counts) 125

7. Diversity Indices 129
7.1 General Description 129
7.2 Analytical Procedures 130
7.3 Diversity of Samples from the Clinch River 134
7.3.1 Species Diversity 134
7.3.2 Hierarchical Diversity 136

8. Summary and Discussion 148
8.1 Introduction 148
8.2 Nature of the Ecosystems from which Data Bases were Selected 148

CONTENTS (continued)

Page

8.3 Methods 149
8.3.1 Relationships Between Methods 149
8.3.2 Data 150
8.3.2.1 Presence-Absence Data 150
8.3.2.2 Quantitative Data 151
8.3.3 Cluster Analysis 151
8.3.4 Ordination 153
8.3.5 Species Diversity and Hierarchical Diversity 153

References 154

LIST OF FIGURES

Number		Page
1	Flow diagram illustrating the computerized information system (CIS) used for TVA's nonfisheries biological data	7
2	Coding form used for keypunching field and laboratory limnological data	8
3	One-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, nine iterations (stress = 0.230)	36
4	Two-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, one iteration (stress = 0.097)	37
5	Two-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, 20 iterations (stress = 0.051)	37
6	Three-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, 43 iterations (stress = 0.001)	39
7	Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients (Table 10) showing faunal similarities between hypothetical samples in Table 9	40
8	Map of the Clinch River in Virginia and Tennessee showing the locations of stations sampled during the 1970 zoomacrobenthic survey	43
9	Stream discharge of the Clinch River at the United States Geological Survey gauging station at Cleveland, Virginia, June-August 1970	45
10	Location of stations in the vicinity of a power plant on the Cumberland River	46
11	Dendrogram computed from cluster analysis of a matrix of coefficients of cophenetic correlation showing similarity between the various correlation and similarity matrices in Tables 17 and 20 (r_{cc} = 0.914)	56

LIST OF FIGURES (Cont.)

Number		Page
12	Dendrogram computed from Q-mode cluster analysis of a matrix of Jaccard's coefficients showing faunal similarities between samples collected from the Clinch River in 1970; data include total insect fauna	58
13	Dendrogram computed from Q-mode cluster analysis of a matrix of Jaccard's coefficients showing faunal similarities between 36 zoomacrobenthic samples collected from stations 4, 7, 8, 9, 10, and 11 in the Clinch River, 1970 ($r_{cc} = 0.93$)	61
14	Dendrogram computed from Q-mode cluster analysis of a matrix of correlation coefficients computed from proportions of each phi size in the substrate after arcsine transformation; shows similarity of substrate between samples collected from the Cumberland River in 1973 ($r_{cc} = 0.760$)	63
15	Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from proportions of each phi size in the substrate after arcsine transformation; shows similarity of substrate between samples collected from the Cumberland River in 1973	64
16	Dendrogram computed from Q-mode cluster analysis of a matrix of Jaccard's coefficients showing faunal similarities between samples collected from the Cumberland River in 1973 ($r_{cc} = 0.852$)	66
17	Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from data transformed with the square-root transformation; shows faunal similarities between samples collected from the Cumberland River in 1973 ($r_{cc} = 0.833$)	67
18	Dendrogram computed from Q-mode cluster analysis of a matrix of correlation coefficients computed from data transformed with the square-root transformation; shows faunal similarities between samples collected from the Cumberland River in 1973 ($r_{cc} = 0.979$)	68
19	Dendrogram computed from Q-mode cluster analysis of a maxtrix of distance coefficients computed from data that had been transformed by the square-root transformation and standardized by rows; shows faunal similarities between samples collected from the Cumberland River in 1975	71

LIST OF FIGURES (Cont.)

Number		Page
20	Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from data that has been transformed by the square-root transformation and standardized by rows; shows faunal similarities between samples collected form the Cumberland River in 1975	72
21	A truncated log-normal distribution fitted to a distribution of species in an aquatic ecosystem not adversely affected by environmental stress	74
22	Response of organism to severe organic enrichment: changes in types of organisms present, population densities, and biological diversity	75
23	Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from stations on the Clinch River unaffected by low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.97)	77
24	Dendrogram computed from R-mode cluster analysis of a matrix of simple matching coefficients, showing distributional similarities of taxa collected from stations on the Clinch River unaffected by low-ph stress that resulted from the 1970 spill of acid (r_{cc} = 0.91)	78
25	Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation; shows distributional similarities of taxa collected from stations on the Clinch River unaffected by low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.92)	79
26	Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from stations on the Clinch River affected by low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.97)	80
27	Dendrogram computed from R-mode custer analysis of a matrix of simple matching coefficients, showing distributional similarities of taxa collected from stations on the Clinch River affected by low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.84)	81

LIST OF FIGURES (Cont.)

Number Page

28 Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation; shows distributional similarities of taxa collected from stations on the Clinch River affected by low-pH stress that resulted from the 1970 spill of acid ($r_{cc} = 0.97$) 82

29 Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from stations on the Clinch River both affected and unaffected by the low-pH stress that resulted from the 1970 spill of acid ($r_{cc} = 0.95$). Only those taxa are included that comprise 10 percent or more of the total sample 83

30 Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation; shows distributional similarities of taxa collected from stations on the Clinch River both affected and unaffected by the low-ph stress that resulted from the 1970 spill of acid ($r_{cc} = 0.97$). Only those taxa that comprise 10 percent or more of the total sample are included 84

31 Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from the Cumberland River in 1973 86

32 Dendrogram computed from R-mode cluster analysis of a matrix of simple matching coefficients, showing distributional similarities of taxa collected from the Cumberland River in 1973 88

33 Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients with data transformed with the square-root transformation and standardized by rows, showing distributional similarities of taxa collected from the Cumberland River in 1973 89

34 Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients with data transformed by the square-root transformation, showing distributional similarities of taxa collected from the Cumberland River in 1973 90

LIST OF FIGURES (Cont.)

Number		Page
35	Dendrogram computed R-mode cluster analysis of a matrix of correlation coefficients with data transformed by the square-root transformation and standardized by rows, showing distributional similarities of taxa collected from the Cumberland River in 1973	91
36	Three-dimensional ordination by nonmetric multidimensional scaling computed from distance coefficients based on presence-absence data, showing faunal similarities between samples collected from the Clinch River in 1970	126
37	Three-dimensional ordination by nonmetric multidimensional scaling computed from distance coefficients based on counts of species, showing faunal similarities between samples collected from the Clinch River in 1970 .	127
38	Species diversity (Brillouin's H) of 36 samples from the Clinch River, 1970; contour interval 0.4	135
39	Species diversity (approximate index H") of 36 zoomacrobenthic samples from the Clinch River, 1970	137

LIST OF TABLES

Number		Page
1	Numerical codes for rivers and streams in the Tennessee Valley, and sources of organisms for bioassays	14
2.0	Codes for identifying methods of limnological sampling and types of gear	15
2.1	Codes for types of sampling equipment and sampling material used in artificial-substrate sampling	16
2.2	Codes for types of sampling equipment and sampling material used in natural substrate removal and organism removal . .	17
2.3	Codes for types of sampling equipment used with emergence traps and volume samplers	18
3	Codes for identifying location of limnological samples on a river or reservoir transect	19
4	Alphabetic codes identifying site (or project) at which sample was collected	20
5	Codes for identifying type of sample (community, parameter, test)	21
6	Codes used to report data units	22
7	Codes for identifying basic type of habitat from which sample was collected or hardness at which bioassay was performed	29
8	Codes for recording instar or size of organisms collected or used in bioassays	30
9	Hypothetical data showing proportions of 10 species at 12 stations	47
10	Matrix of distance coefficients computed from the data in Table 9 after arcsine transformation	48
11	Generalized trophic, functional classification of zoomacrobenthic invertebrates	49
12	Hierarchical classification of the trophic-functional role of organisms; includes only those categories that occurred in samples	50

LIST OF TABLES (Cont.)

Number		Page
13	Hierarchical classification assigned for zoomacrobenthic invertebrates based on functional morphology: head position, body shape, and respiratory organs	51
14	Range and mean of physicochemical data for the Clinch River, June to September 1970	52
15	Descriptions of habitats at stations 4, 7, 8, 9, 10, and 11 on the Clinch River, 1970	53
16	Coefficients of correlation, distance, and similarity: abbreviations, equations, and upper and lower limits . . .	92
17	Contingency table (2 X 2) defining the terms a, b, c, and d, as used in the equations in Table 13	94
18	Effect of the transformation log $(X_{ij} + 1)$, where X_{ij} is the abundance of the ith species in the jth sample . .	94
19	Effect of the transformation $\sqrt{(X_{ij} + 0.5)}$, where X_{ij} is the abundance of the ith species in the jth sample . .	94
20	Labels of correlation and distance matrices with various transformations	95
21	Matrix of coefficients of cophenetic correlation computed between corresponding elements of 21 correlation and similarity matrices	96
22	Coefficients of cophenetic correlation comparing distance matrices and selected correlation and similarity matrices	98
23	Number of taxa in each major taxonomic group in the Clinch River (1970) before and after relative species abundance was determined and rare taxa were eliminated . .	99
24	Twenty-nine taxa and their respective trophic codes for the reduced Clinch River data set, 1970	100
25	Number of taxa with a relative abundance ≥ 0.01 divided by the total number of taxa per station, Clinch River, 1970	101
26	Taxa with a relative abundance ≥ 0.01 as percent of the total number of taxa per station, Clinch River, 1970 . . .	101
27	Total number of organisms per station, Clinch River, 1970	102

LIST OF TABLES (Cont.)

Number Page

28 Twenty-nine taxa with relative abundance $\geq$0.01 as percent of total number of organisms per station, Clinch River, 1970 102

29 Analytical technique, type of comparison, and the number and type of similarity coefficients used to analyze the reduced (1970) Clinch River data set 103

30 Cophenetic correlation values (r_{cc}) for 24 dendrograms computed from the Clinch River data set, June to August 1970 104

31 Results of Q-mode cluster analysis of zoomacrobenthic samples from the Clinch River, 1970; minimum level of similarity used to define clusters = 0.62; Jaccard's coefficient 105

32 Results of Q-mode cluster analysis of zoomacrobenthic samples from the Clinch River, 1970; level of similarity used to define clusters = 0.06; correlation coefficient 6, $\sqrt{Y + 0.5}$ transformation and standardization by rows 106

33 Results of Q-mode cluster analysis of zoomacrobenthic samples from the Clinch River, 1970; level of similarity used to define cluster = 1.2; distance coefficient 6, $\sqrt{Y + 0.5}$ transformation and standardization by rows . . 107

34 Results of Q-mode cluster analysis of zoomacrobenthic samples from the Clinch River, 1970; level of similarity used to define clusters = 4.8; distance coefficient 7, $\sqrt{Y + 0.5}$ transformation and standardization by rows 108

35 Summary of results of Q-mode cluster analyses, Cumberland River, 1973 109

36 Summary of results of Q-mode cluster analyses, Cumberland River, 1975 111

37 Results of R-mode cluster analysis of Jaccard's (S_j) and simple-matching (S_{sm}) coefficients, Clinch River data set, 1970 113

38 Results of R-mode cluster analysis of distance coefficients, Clinch River data set, 1970 114

39 Results of R-mode cluster analysis of S_{sm} and S_j, Clinch River stations unaffected by the 1970 pH stress . . 116

LIST OF TABLES (Cont.)

Number		Page
40	Trophic-functional codes for taxa clustered by R-mode cluster analysis of Jaccard's (S_j) and simple-matching (S_{sm}) coefficients, Clinch River data set, 1970	117
41	Trophic-functional codes for taxa cluster by R-mode cluster analysis of Dist 7 coefficients Clinch River data set, 1970	118
42	Trophic-functional codes for taxa clustered by R-mode cluster analysis of simple-matching (S_{sm}) and Jaccard's (S_j) coefficients, Clinch River, 1970	119
43	Results of R-mode cluster analysis of S_j and S_{sm} after reordering clusters according to trophic-functional codes, Clinch River, 1970	120
44	Results of R-mode cluster analysis of Dist 7 coefficients after reordering clusters according to trophic-functional codes, Clinch River, 1970	122
45	Clusters of taxa defined by R-mode cluster analysis of Jaccard's and the simple matching coefficients, Clinch River, 1970	123
46	Species diversity (Brillouin's H) of 36 zoomacrobenthic samples from the Clinch River, 1970	140
47	Species diversity (approximate index H") of 36 zoomacrobenthic samples from the Clinch River, 1970	141
48	Hierarchical taxonomic diversity (Brillouin's H) of zoomacrobenthic samples from the Clinch River, 1970	142
49	Component of species diversity (H) at each level in the trophic-functional hierarchy for five samples or subsamples collected immediately after the acid spill on the Clinch River, 1970	145
50	Percent of species diversity (H) contributed at each level in the trophic-functional hierarchy for five samples or subsamples collected immediately after the acid spill on the Clinch River, 1970	146
51	Component of species diversity (H) at each level in the head-body-respiratory functional morphology hierarchy for five samples or subsamples collected immediately after the acid spill on the Clinch River, 1970	147

LIST OF ABBREVIATIONS AND SYMBOLS

CIS	--	computerized information system
Corr	--	correlation coefficient; for use with species counts
ΔT	--	temperature change, °C
$\bar{d}$	--	Wilhm-Dorris diversity index
Dist	--	distance coefficient; for use with species counts
DM/IS	--	data management/information system
DMS	--	data management system
H	--	Brillouin's index
MDS	--	multidimensional scaling
NT	--	number of taxa
NTSYS	--	Numerical Taxonomy System
Q-mode comparison	--	pairwise comparison between columns; used to determine similarity between sites or stations on basis of biotic assemblages present
r_{cc}	--	coefficient of cophenetic correlation
R-mode comparison	--	pairwise comparison between rows; used to determine similarity between species assemblages on basis of distribution among samples
SAS	--	Statistical Analysis System
SCE	--	standing crop estimates
S_J	--	Jaccard's coefficient; for use with presence absence data
S_{SM}	--	simple matching coefficient; for use with presence-absence data
TF	--	Trophic-Functional code
7LB&MS	--	Station 7, left bank and midstream
7RB	--	Station 7, right bank
TSO	--	time-sharing option
$\geq$	--	greater than or equal to
$<$	--	less than

ACKNOWLEDGMENTS

The authors express their appreciation to Dr. Ralph H. Brooks, Mr. Billy G. Isom, and Dr. Harrison Hickey, Tennessee Valley Authority, for their cooperation and support; Mr. Clinton W. Hall, Project Officer, Environmental Protection Agency, for his patience and support; and Ms. Rachel C. Strong for her drafting work. Acknowledgment is also given to Dr. Brian Armitage, Mr. Clay Barr, and Ms. Sandra Emond of the Tennessee Valley Authority, for their help in developing many of the coding schemes for the computerized information system. Recognition is also given Dr. Cornelius Weber, Environmental Protection Agency, and Mr. Tom Toole and Dr. Ken Tennessen, Tennessee Valley Authority, for their assistance in updating the BIO-STORET species list.

SECTION 1

INTRODUCTION

An essential part of achieving national self-sufficiency in energy is minimizing the adverse environmental impacts that may accompany accelerated development and increased use of energy resources. The Tennessee Valley Authority (TVA) has been studying ways to evaluate and minimize these impacts through the Federal Energy/Environment Research and Development Program, coordinated by the Office of Energy, Minerals, and Industry (OEMI) of the Environmental Protection Agency (EPA). This program is designed to (1) add environmental objectivity and balance to the mission of the Department of Energy, (2) prevent delays in development of energy resources that are caused by inadequate environmental information, (3) develop cost-effective strategies for pollution control, (4) promote transfer of energy-related environmental information, and (5) project the impacts of future energy technologies (Environmental Protection Agency 1976).

As part of the federal interagency agreement with EPA, TVA's applied research program undertook a comprehensive evaluation of the impacts of energy-related technologies on the aquatic environment. The research discussed in this milestone report summarizes the work completed during the first two years of the project on Task 1 (Information Systems Development) of Subagreement 10 (Consolidation of Thermal Impacts on Freshwater Shellfish, Insects, and Other Biota). The overall objective of Task 1 was to develop the capability to measure and evaluate existing and expected environmental impacts of energy-related technologies on important biotic assemblages (nonfish) in the aquatic environment.

To accomplish this objective, the first priority was to develop a computerized information system (CIS). This system was designed to be inexpensive to operate, user oriented, adaptable for use in both routine monitoring studies and research projects, capable of performing a variety of analytical procedures, and able to interface with EPA's BIO-STORET system.

Using these criteria and the Statistical Analysis System (SAS) (Barr et al., 1976) as the basic data management system, a CIS was developed that could accommodate 19 different biotic and abiotic parameters. Since it was important to conserve space while including information for each parameter, a series of numeric and alphanumeric codes was developed. These codes not only reduced storage requirements, but saved time and resources each time the data were sorted, compiled, and analyzed. Additional software packages were also adapted or developed to interface with the CIS. These routines expanded the analytical and computer-graphics capabilities of the system.

After the CIS had been developed, the next priority was to evaluate three procedures used to analyze data from biological surveys. The first procedure was the cluster analysis routine in the Numerical Taxonomy System (NTSYS) written by F. J. Rohlf and associates. Two applications of cluster analysis were considered. 1. Q-mode analysis was used to determine the similarity of different samples or sampling stations on the basis of the species found in each sample. 2. R-mode analysis was used to identify associations of species on the basis of their spatial and temporal distributions.

The second analytical technique tested was nonmetric multidimensional scaling (MDS), in which the information was presented in a scatter diagram that was examined without first assuming that clusters were present. MDS was examined as, and might be considered, an alternative to cluster analysis for determining whether species form distinct biological assemblages.

The third analytical procedure considered was hierarchical diversity analysis. This technique was an extension of the diversity measures commonly used to summarize biological data from environmental surveys. It was used to determine (1) the usefulness of two trophic-functional coding schemes developed for zoomacrobenthic organisms and (2) the importance of species diversity compared to diversities at higher taxonomic levels.

SECTION 2

CONCLUSIONS AND RECOMMENDATIONS

2.1 INTRODUCTION

The primary purpose of "Task 1: Information Systems Development" was to develop a system for storing, retrieving, and analyzing data on aquatic ecosystems. A secondary objective was to examine and evaluate forms of various analytical techniques routinely used to summarize and predict the environmental impacts of accelerated development of energy resources. This report summarizes the system and provides tests and demonstrations of quantitative procedures used to analyze large data sets. The methods tested are cluster analysis, ordination, species diversity, and hierarchical diversity analysis.

2.2 GENERIC

Throughout the study the paramount importance of pertinent and representative data in assuring sound environmental interpretations was evident. Although this point is intuitive, it is often a prerequisite that is not achieved. One controllable and important part of assuring that sound data will be available is to make sure that the right questions are asked. Attempts at general-purpose monitoring should be abandoned, and statistically trained ecologists familiar with sampling theory should be brought into study teams from the very start. Moreover, extensive preliminary sampling should be undertaken to discover subtleties of the ecosystem. During actual monitoring unessential parts of the preliminary sampling plan can be abandoned in favor of more detailed sampling of critical areas. Cursory study of a few adequate samples will provide greater insight than detailed examination of many inadequate ones. Samples must not only be well located but also of sufficient size that possible errors of sampling are minimized. For example, with small sample sizes the likelihood that rare species may be missed by chance alone increases. Thus, species may be present in some samples and absent from others due to chance and not due to environmental reasons.

Biological data may be recorded as presence-absence data or quantitative data. Presence-absence data may be obtained much more quickly and less expensively, but information provided by differential abundances is lost. For most purposes of applied aquatic ecology, quantitative data are preferred, but more work needs to be done to assess the utility of presence-absence data. Such data are often all that is available, and quick results may be essential in times of acute environmental crisis.

2.3 CLUSTER ANALYSIS

Cluster analysis is a multivariate quantitative procedure that provides a classification of samples. The method is often useful in environmental surveys because it provides readily interpretable results in the form of a tree-like graph called a dendrogram. Similarities among all samples in a study are presented simultaneously in the dendrogram, although a measurable amount of distortion may be introduced during the clustering procedure. Fortunately, this distortion is measurable, and interpretation of dendrograms with unacceptable levels of distortion can be avoided.

In Q-mode analyses, samples with similar faunas are grouped together. In aquatic surveys involving a perturbation of the environment, one expects samples from upstream control stations to be grouped with downstream controls in the zone of recovery. Unfortunately, faunas at downstream stations are sometimes inherently different from upstream ones because of changes in stream gradient, discharge, substrate, and other factors.

R-mode cluster analyses indicate similarities among, or associations of, species on the basis of their distribution and on the basis of abundance if quantitative coefficients are used. Q-mode analyses have been used much more frequently than R-mode in applied ecology, but R-mode analysis has promise as a method of comparing faunal associations from stream to stream or basin to basin.

A further disadvantage of cluster analysis and a further source of distortion is that it forces samples into clusters whether or not such clusters exist in nature. Although the total amount of distortion can be measured, its effects cannot be determined precisely without time-consuming study of the raw data.

For cluster analysis of presence-absence data, we recommend the use of Jaccard's coefficient, which bases similarity between stations on only the mutual presence of organisms at stations and not on their absences. Quantitative data should be analyzed with correlation or distance coefficients after the data have been transformed with a square-root transformation to reduce inordinate effects of highly abundant species. Future research should be directed toward evaluating use of presence-absence data.

2.4 ORDINATION

The results of ordination are presented as a scatter diagram in which stations are plotted as points in a space defined by axes that represent the faunal similarities of stations to each other. The goal of ordination is similar to that of Q-mode cluster analysis because both are computed from similarities among stations. It differs in that no a priori assumptions need be made about the presence of clusters in the data. Although little use has been made of ordination in applied aquatic ecology, it has potential as an analytical tool. Streams are linear ecosystems, and to cluster stations along such an environmental gradient may require forcing them into unrealistic configurations. An ordination can let the investigator determine if clusters exist.

Several methods of ordination are available. We have tested only nonmetric multidimensional scaling, a technique that seems ideally suited to data from aquatic environmental surveys. Because little use has been made of ordination, we have evaluated it only against the results of cluster analysis. In general, it seems to give comparable results, being especially sensitive to inadequate data, from which it produces uninterpretable results. Future research should stress additional study of nonmetric multidimensional scaling, especially applications to presence-absence data.

2.5 INDICES OF DIVERSITY

An index of species diversity is a single statistic that expresses both the number of species present and the evenness of distribution of organisms among species. Unlike cluster analysis and ordination, it does not consider which species are present. Thus, a sample collected in a zone of recovery downstream from a source of pollutional stress may have a species diversity equal to samples from upstream control stations, yet the species may be altogether different. Because cluster analysis and ordination measure a different aspect of community structure from species diversity, the methods should be used in combination for best results.

The concept of species diversity has been misunderstood, misapplied, and subsequently much maligned in the ecological literature. Too many investigators have overlooked the nonuniqueness of the relationship between a particular community structure and the index of species diversity computed from it. Others have sought global values of species diversity to indicate healthy or damaged ecosystems, overlooking the dependence of all such indices on sample size. Nevertheless, species diversity is a useful tool for applied aquatic ecologists. If used properly and not overinterpreted, it gives a useful and efficient measure of community structure for communicating information about the state of the ecosystem to nonbiologists.

Species diversity may be partitioned according to categories in the taxonomic hierarchy to give diversity components contributed by orders, families, genera, and species. For some purposes of applied aquatic ecology, discrimination of genera may provide as much environmental information as identification of species, with savings of time and money. Diversity may also be partitioned according to other hierarchical classifications that consider functional morphology and feeding strategies. Hierarchical diversity was tested in all three ways in the report and was found to be useful.

Species diversity can be computed with a number of equations. We recommend the use of Brillouin's equation from information theory. Other equations from information theory are biased and often give misleading results. Hierarchical diversity analysis should be applied in future studies to help reduce the cost of aquatic environmental surveys.

2.6 FINAL STATEMENT

Although most of the methods of applied aquatic ecology were derived from theoretical ecology, the goals of the two sciences are not the same. It is incumbent on applied ecologists to adapt and modify existing methods to suit their needs and to address questions of urgency. At the same time, they must continue to recognize their dependence on the groundwork laid by more theoretically inclined ecologists.

SECTION 3

DEVELOPMENT OF THE INFORMATION SYSTEM

3.1 GENERAL DESCRIPTION OF THE INFORMATION SYSTEM

When TVA initiated the research program on effects of energy use and development, Task 1 (Information Systems Development) was responsible for developing a system through which to measure and evaluate the impact of energy technologies on biotic assemblages. To accomplish this task, a computerized information system (CIS) was developed to accommodate biological data. Criteria for design of the system were that it must interface with EPA's BIO-STORET system, be inexpensive to operate, be user oriented, be adaptable for use with both routine monitoring and research programs, and perform a variety of analytical functions.

A flow diagram illustrating the CIS and steps involved in processing the data is presented in Figure 1. In its present stage of development, the CIS uses (1) the Statistical Analysis System (SAS) (Barr et al., 1976) to create and manipulate data sets; print, sort, rank, and store data; and perform analyses such as simple descriptive statistics, Duncan's multiple range test, analysis of variance, correlation, prohibit analysis, and regression; (2) the Numerical Taxonomy System (NTSYS) for cluster analysis and ordination with nonmetric multidimensional scaling; (3) the MIT-SNAP programs (Hoaglin and Welsch, 1975) for resistant regression and box plotting; (4) user-written FORTRAN programs to calculate diversity indices; and (5) Tektronix software for producing instantaneous hard-copy graphics.

3.2 PREPARATION OF DATA

To store and manipulate the large volume of information available within TVA, a series of numeric and alphanumeric codes was developed to allow for easier storage and retrieval of biological data; to save time and resources each time the data were sorted, compiled, and analyzed; and to centralize storage of environmental data.

An example of the standard coding form used to transfer field and laboratory data into a format for keypunching is presented in Figure 2. The type of program (macrobenthos, phytoplankton, or zooplankton), account number, job number, data originator, address and phone number of originator, and sheet number are recorded at the top of each sheet.

The standard coding form has 80 columns and 26 rows. The 80 columns are divided into 19 data fields, with 1 to 16 columns per field. The data fields correspond to the following variables:

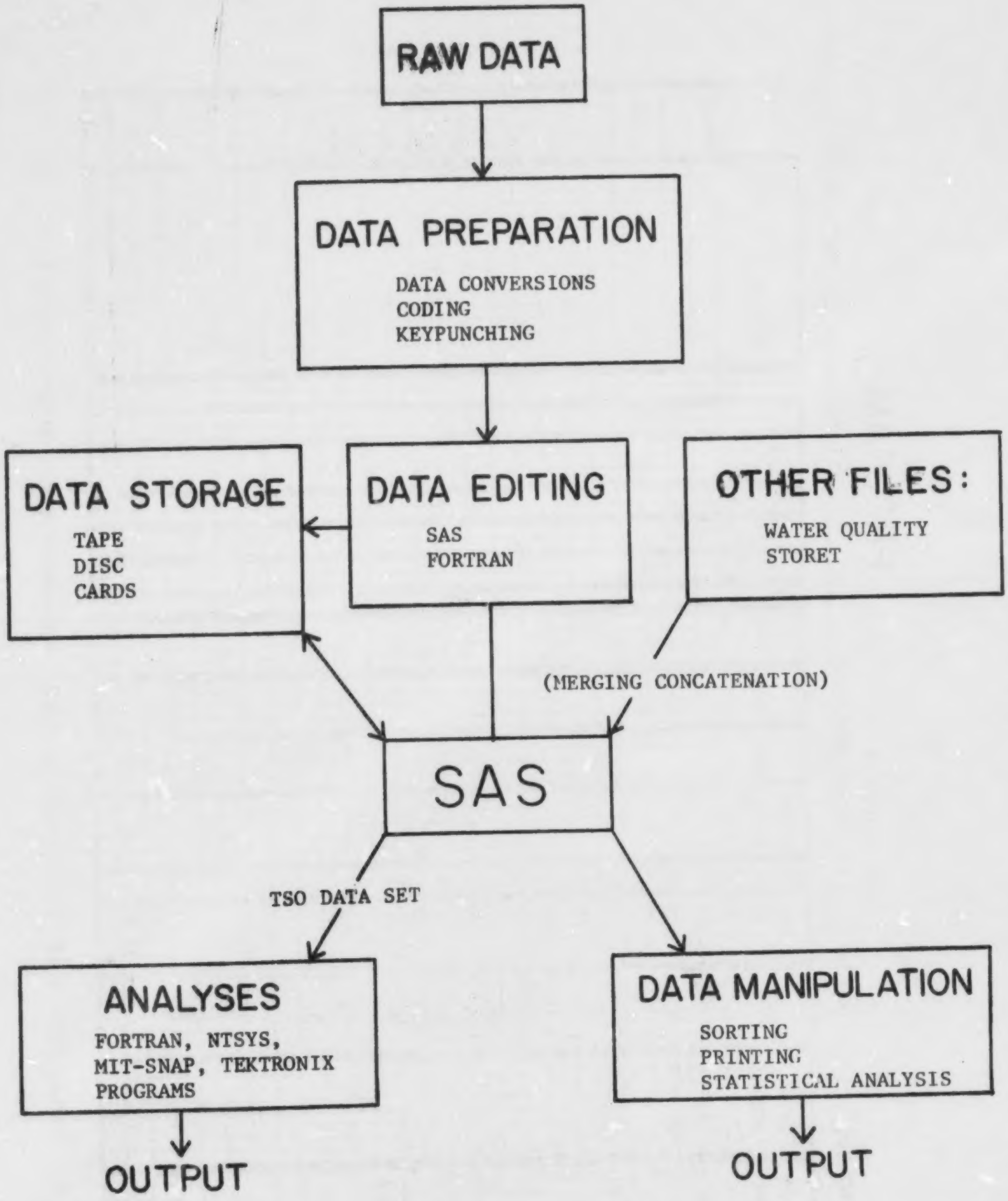
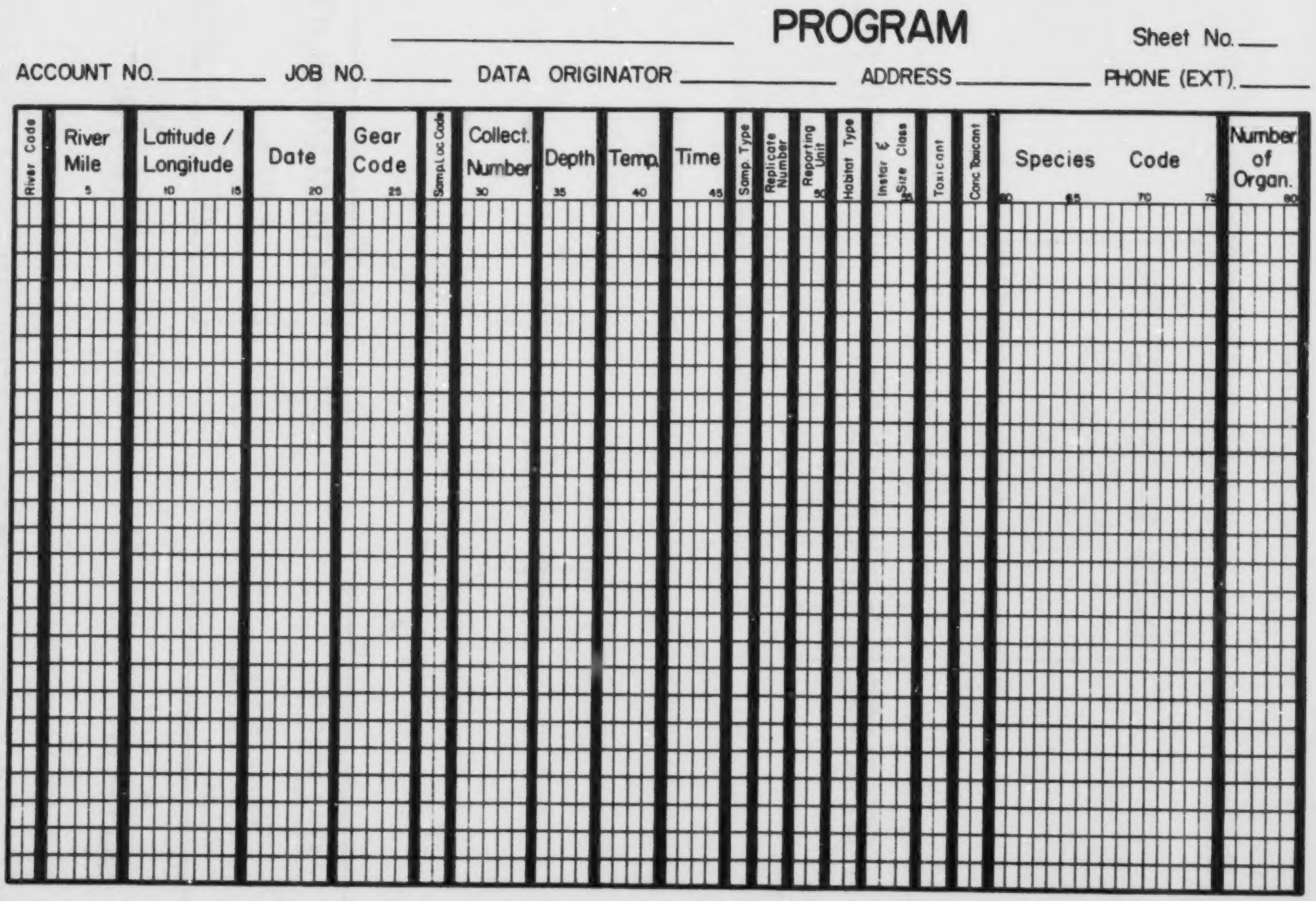

RAW DATA
DATA PREPARATION
DATA CONVERSIONS
CODING
KEYPUNCHING
DATA STORAGE
TAPE
DISC
CARDS
DATA EDITING
SAS
FORTRAN
OTHER FILES:
WATER QUALITY
STORET
(MERGING CONCATENATION)
SAS
TSO DATA SET
ANALYSES
FORTRAN, NTSYS,
MIT-SNAP, TEKTRONIX
PROGRAMS
OUTPUT
DATA MANIPULATION
SORTING
PRINTING
STATISTICAL ANALYSIS
OUTPUT

Figure 1. Flow diagram illustrating the computerized information system (CIS) used for TVA's nonfisheries biological data.



______________________ PROGRAM Sheet No.___

ACCOUNT NO.________ JOB NO.______ DATA ORIGINATOR ____________ ADDRESS __________ PHONE (EXT).______

River Code	River Mile 5	Latitude / Longitude 10 15	Date 20	Gear Code 25	Sampl. oc Code	Collect. Number 30	Depth 35	Temp. 40	Time 45	Samp. Type	Replicate Number	Reporting Unit 50	Habitat Type	Instar & Size Class 55	Toxicant	Conc Toxicant	Species Code 60 65 70 75	Number of Organ. 80

Figure 2. Coding form used for keypunching field and laboratory limnological data.

1. river code
2. river mile
3. latitude and longitude
4. date
5. gear code
6. sample location code
7. collection number
8. depth
9. temperature
10. time
11. sample type
12. replicate number
13. reporting unit
14. habitat type
15. instar and size class
16. toxicant
17. concentration of toxicant
18. species code
19. number of organisms

The first data field consists of a two-digit numeric code for rivers and streams in the Tennessee Valley (Table 1) and sources of cultures utilized in bioassays.

Columns 3 to 7 contain the river-mile location of a sampling site. This information specifies the distance a site is located upstream from the mouth of the river and is usually obtained from a basin navigation map.

The latitude and longitude of a site are entered in columns 8 to 15. These data are reported in degrees and minutes. The first four columns are for latitude, and the second four columns are for longitude. For example, a sample taken at latitude 36°19' and longitude 86°23' is be entered as "36198623."

Columns 16 to 21 specify the date of sampling or bioassay. The date is entered numerically by year, month, and day with "730615" denoting June 15, 1973.

The code identifying the type of equipment used to collect a sample, is entered in columns 22 to 26. This hierarchical gear code has six general categories: undefined, artificial substrates, natural substrate removal, direct organism removal, emergence traps, and volume samplers (Table 2). Five of these categories are divided into subgroupings to accommodate information such as sampler name, mesh size, or type of substratum used. Each sampling method or device has a five-character alphanumeric code, starting with the letter "A" and followed by four numbers. Generally, the first two digits identify the type of sampling equipment, and the next two digits describe the specific type of sampling material used in artificial substrate samplers (Tables 2.1-2.3). For example, the four major categories of artificial substrate samplers are baskets, trays, flat surfaces, and sterile indigenous substrates; and one or more substrate materials can be used with each. Thus, the code "A1210" denotes a tray sampler with a rock substratum. The code "A1410" indicates that the sampler uses sterile, indigenous rocks (Table 2.1).

Table 2.2 lists the gear codes for natural substrate removal and organism removal. The gear codes for natural substrate removal have five main categories, three of which give specific information on the name or type of sampler used. For example, the code "A2200" is the general code number for corers, but "A2250" is the code for the Benthos 2170 gravity corer. Again, the last two digits indicate the type of gear used.

Columns 27 and 28 contain a two-digit code (Table 3) that identifies the distance a sample was collected from the riverbank. The codes "01" to "49" indicate distance from the right bank (facing downstream), and the codes "50" to "98" are coded distances from the left bank. Code "63," for example, indicates that a sample was taken 25.1 to 30.0 m from the left bank, facing downstream.

The collection number is a five-character alphanumeric code in columns 29 through 33. The first two characters identify the location or project at which at which a sample was collected. The remaining three digits refer to the 1st, 2nd, 3rd, . . . , or 999th time a sample was collected at the site. Table 4 lists only the alphabetic prefix of the collection code.

The depth (in meters) a sample was collected is recorded in columns 34 to 37. For example, samples collected 3 and 24.5 m below the surface are recorded as "03.0" and "24.5." The decimal points occupies a separate column. If the depth occupies only three columns, a zero is placed on the left, as in "03.0." A computer printout of 3.0 is obtained by right justifying all integers.

The water temperature (in °C) at which the sample was collected or at which the bioassay was performed is entered in columns 38 to 41. A temperature of 21.2°C is entered as "21.2." Again, the decimal point is entered in a separate column. If the temperature occupies only three columns, zeros are placed to the left to fill the space; e.g., 9.6°C is entered as "09.6."

The time of day (0000 to 2400) a sample was collected, or the length of time for a bioassay, is recorded in columns 42 through 45. A sample collected at 2:45 p.m. is entered as "1445." A data entry for hour 48 of a bioassay is entered as 0048.

The code for sample type is a one-character alphabetic code entered in column 46 (Table 5). This code identifies the type (community, parameter) of sample or data collected or reported, such as periphyton, phytoplankton, chlorophyll, or bioassay.

The replicate number is entered in columns 47 and 48. Because replicate sampling and testing is done for most surveys and bioassays, it is important to know not only how many replicates were collected but the replicate one is dealing with. This code refers to a given sample or replicate and not to the total number collected. Thus, "06" refers to the sixth sample collected at a particular sample station, and "10" refers to the tenth sample taken at the same station.

The units in which the data are reported are coded as a two-character alphanumeric code entered in columns 49 and 50. Table 6 lists the units commonly used. These units are listed according to general categories of area, chlorophyll pigments, percentages, productivity-respiration, radiation-light, rate, flags and foul-ups, length, number, ratio, temperature, time, turbidity, volume, weight, and zooplankton.

The code for habitat type or ecologic zone is entered in columns 51 and 52. This two-digit code identifies the habitat (ecologic zone and dominant substratum) found at each site. In the case of bioassays, the code indicates the range of hardness (Table 7).

Columns 53 to 55 contain a three-character code for instar and size class. The first column refers to the stage of development or to the set of characters used to express the stage of development of a specimen:

I--instar
S--general size class
L--length class
P--pupal stage
H--head capsule
A--immature

The next two columns contain a numeric code that represents size intervals in millimeters (Table 8). Because instars are not always determined by measurements of total body length, it is important to specify on the laboratory bench sheet whether head capsule dimensions, wing pad length, or other measurements were used to determine the instar. Examples of common codes are:

L48--length class: 100.0 to 125.0 mm
S07--size class: 0.61 to 0.70 mm
P00--pupal stage (interval code disregarded)
A00--nauplius immature (interval code disregarded)
H03--head capsule width: 0.21 to 0.30 mm

Additional bioassay information where appropriate is recorded in columns 56 through 59. Symbols from the periodic table or a coded listing of compounds are entered in columns 56 and 57, and the concentration of the toxicant is entered in columns 58 and 59. If data on a toxicant are recorded, the units code entered in columns 49 and 50 gives the units of the concentration of the toxicant. For example, a combination of "68" in columns 49 and 50 (units; see Table 6), "HG" in columns 56 and 57 (toxicant), and "05" in columns 58 and 59 (concentration of toxicant) indicates that a concentration of 5 µg/l mercury was used in the bioassay.

The next field refers to the 16-unit biological species code, which is entered in columns 60 to 75. This code identifies eight taxonomic categories. Columns 60 and 61 identify the phylum or division; columns 62 and 63 identify class; columns 64 and 65 identify order; columns 66 and 67 identify family; columns 68 and 79 identify genus; columns 71 to 73 identify species; and columns 74 and 75 identify either the variety, form, or authority. The numeric values are 01 to 99 for the two-digit fields and 001 to 999 for the three-digit fields.

A typical 16-unit species code is "1801190100100100," which gives the following information:

18	Phylum:	Arthropoda
1801	Class:	Crustacea
180119	Order:	Branchiura
18011901	Family:	Thalestridae
18011901001	Genus:	*Argulus* sp.
18011901001001	Species:	*japonicus*
1801190100100100	Authority:	Thiele

The codes for species found in the Tennessee Valley were compiled in the Synoptic Catalog of Algae and Aquatic Invertebrates for the Tennessee Valley included as Appendix A of this report.

The total number of organisms per data entry (row) is recorded in columns 76 to 80. For field collections this number refers to the number of organisms collected per data entry (sample or replicate). To indicate that 97 individuals of *Chironomus tentans* Fabricius were found per square meter, "NG" is entered in the units field (columns 49 and 50), "1802111501602000" is entered as the species code (columns 60 to 75), and "00097" is entered as the number of organisms (columns 76 to 80). For recording bioassay results, columns 76 to 80 are utilized to indicate the number of organisms killed and the total number of organisms tested, respectively. For example, to indicate that after 48 hours of an acute bioassay, 28 of 40 specimens of *C. tentans* had died, "0048" is entered in columns 42-45, "H" is entered in column 46, "1802111501602000" is entered in columns 60-75, "28" is entered in columns 76-77, column 78 is left blank, and "40" is entered in columns 79-80.

3.3 EDITING OF DATA

After the biological data are coded and transcribed onto the coding form, they are keypunched onto cards of a specified color, depending on the type of data collected. For example, zooplankton data are punched on yellow cards, zoomacrobenthic data on red cards, phytoplankton data on green cards, carbon-14 data on orange cards, and chlorophyll data on white cards.

After the data have been punched, they are read into the computer where a SAS program sorts and merges the data with another SAS data set that contains the taxonomic name associated with each species code. If a data record does not have a taxonomic name for a given code, an error message is printed. A list of the data is then checked by the principal investigator to ensure that all other requirements of the data have been met. If a record is incomplete, new data are added by means of a user-written FORTRAN program. For instance, if a zoomacrobenthic species was found in sample replicates 1, 5, 6, and 9 at a site, the FORTRAN program would insert zeros for replicate numbers 2, 3, 4, 7, 8, and 10 if ten replicated samples had been taken at the station. Once this is done, SAS can be used to sort, print, and perform statistical analyses. Additional data can also be merged with the test data set at this time.

Once these steps have been completed, a formatted data set can be output via a time-sharing option (TSO) or batch processed for use with other programs. Programs used on TSO or submitted to batch from TSO are:

1. SPECLIST provides a species list. This list is checked to ensure that all organisms listed were actually found in the study.

2. MATRIX provides a data matrix for use in NTSYS. NTSYS language control cards are inserted on TSO, and the programs are submitted to batch.

3. DIVER calculates the diversity index H'.

4. MIT-SNAP; median polish is used from this software package.

5. NTSYS provides 7 types of clustering, 7 types of ordination, 24 data transformations, and 22 indices of similarity or dissimilarity.

6. User-written programs for species diversity (H) and hierarchical diversity.

TABLE 1. NUMERICAL CODES FOR RIVERS AND STREAMS IN THE TENNESSEE VALLEY, AND SOURCES OF ORGANISMS FOR BIOASSAYS (UTILIZED IN COLUMNS 1 AND 2 OF CODING FORM--FIGURE 2)

Numerical code	Alphabetical code[a]	Identification
00	TRM	Tennessee
01	ERM	Elk
02	HRM	Hiwassee
03	LTRM	Little Tennessee
04	FBRM	French Broad
05	HORM	Holston
06	HSRM	South Fork Holston
07	WRM	Watauga
08	CURM	Cumberland
09	CRM	Clinch
11	SRM	Stones
12	CFRM	Caney Fork
13	MRM	Mississippi
14	OHRM	Ohio
15	GRM	Green
16	DRM	Duck
17	OBRM	Obey
18	BRM	Buffalo
19	PRM	Powell
20	HNRM	North Fork Holston
21	HPRM	Harpeth
22	NRM	Nolichucky
23	PGRM	Pigeon
24	SQRM	Sequatchie
25	CHRM	Cheoah
26	NTRM	Nantahala
27	TGRM	Tuckasegee
28	NLRM	Nottely
29	TCRM	Toccoa
30	EMR	Emory
31	OCO	Ocoee
32	TELL	Tellico
33	CFM	Clear Fork River
34	TCM	Town Creek
35	CCM	Crooked Creek
36	YCM	Yellow Creek
37		Research Station, Browns Ferry
38		Rearing Ponds, NFDC, Muscle Shoals
39		EPA Laboratory, Corvallis, OR

---

[a] Used by other data storage systems in TVA.

TABLE 2.0. CODES FOR IDENTIFYING METHODS OF LIMNOLOGICAL SAMPLING, AND TYPES OF GEAR (UTILIZED IN COLUMNS 22-26 OF CODING FORM--FIGURE 2)

A0000	Undefined
A1000	Artificial substrates
A2000	Natural substrate removal
A3000	Direct organism removal
A4000	Emergence traps
A5000	Volume samplers

TABLE 2.1. CODES FOR TYPES OF SAMPLING EQUIPMENT AND SAMPLING MATERIAL USED IN ARTIFICIAL-SUBSTRATE SAMPLING

A1000	Artificial substrate
A1100	Baskets
A1110	Rock
A1111	Rocks, bagged (basket collected in bag)
A1112	Rocks, unbagged (basket not collected in bag)
A1120	Leaf
A1130	Brush
A1140	Conservation webbing
A1150	Other synthetic material
A1160	Balls (porcelain or other material)
A1170	Concrete and pebble blocks
A1171	Concrete, pebble blocks, and conservation webbing
A1200	Trays
A1210	Rock
A1220	Pebbles
A1230	Sand
A1240	Silt
A1250	Clay
A1260	Mud
A1270	Conservation webbing
A1280	Other synthetic material
A1300	Flat surfaces
A1310	Multiplate samplers
A1320	Glass slides or plates
A1330	Plexiglass slides or plates
A1340	Plastic strips or sheets
A1350	Polyethylene plates
A1360	Polyurethane
A1370	Polystyrene
A1380	Iron plates
A1390	Ceramic tile or block
A13AA	Wood
A13BB	Cement tile or block
A1400	Sterile indigenous substrate
A1410	Rock
A1420	Wood
A1430	Aquatic vascular plants

TABLE 2.2. CODES FOR TYPES OF SAMPLING EQUIPMENT AND SAMPLING MATERIAL USED IN NATURAL SUBSTRATE REMOVAL AND ORGANISM REMOVAL

A2000 Natural substrate removal

A2100	Dredges
A2110	Eckman
A2120	Petersen
A2130	Ponar
A2140	Franklin-Anderson
A2150	Shipek
A2160	Dietz-LaFond
A2170	Orange peel dredge
A2180	Tonolli spiraling mud burrower
A2200	Corers
A2210	Vertical core sampler
A2211	Cork borer
A2212	Dendy inverted sampler
A2213	Phleger corer
A2214	Ewing piston corer
A2220	FRB multiple corer
A2230	Peat corer
A2240	Deep core sampler
A2250	Benthos 2170 gravity corer
A2260	Alpine 211 gravity corer
A2270	PVC pipe corer
A2280	Glass tube corer
A2290	Kajak corer
A2300	Area samplers
A2310	Surber square foot sampler
A2320	Wilding square foot sampler
A2330	Hess circular sampler
A2340	Neill sampler
A2350	Dome sampler
A2360	Diver-Actuated Sampler - Circular
A2361	Diver-Actuated Sampler - Square
A2400	Scrapes
A2500	Ooze suckers

A3000 Organism removal

A3100	Kick nets
A3200	Hand net sweeps
A3300	Drift nets
A3400	Grab samples
A3500	Trawls

TABLE 2.3. CODES FOR TYPES OF SAMPLING EQUIPMENT USED WITH EMERGENCE TRAPS AND VOLUME SAMPLERS

	A4000 Emergence traps
A4100	Light traps
A4110	Lantern-sheet method
A4120	Incandescent
A4130	Fluorescent
A4140	Black light (UV)
A4200	Submerged traps
A4300	Floating
A4400	Aerial net traps
A4500	Staked box traps
A4600	Hand net sweeps
	A5000 Volume samplers
A5100	Juday traps
A5200	Kemmerer bottles
A5300	Van Dorn bottles
A5400	Clarke-Bumpus plankton sampler
A5500	Nansen bottle
A5600	Grab samples
A5700	Undefined
A5800	0.5-Meter tow net (#20 mesh)

TABLE 3. CODES FOR IDENTIFYING LOCATION OF LIMNOLOGICAL SAMPLES ON A GIVEN RIVER OR RESERVOIR TRANSECT (UTILIZED IN COLUMNS 27 AND 28 OF CODING FORM--FIGURE 2)

Code	Interval[a] (meters)	Code	Interval[b] (m)
01	0.1 - 1.0	50	0.1 - 1.0
02	1.1 - 2.0	51	1.1 - 2.0
03	2.1 - 3.0	52	2.1 - 3.0
04	3.1 - 4.0	53	3.1 - 4.0
05	4.1 - 5.0	54	4.1 - 5.0
06	5.1 - 6.0	55	5.1 - 6.0
07	6.1 - 7.0	56	6.1 - 7.0
08	7.1 - 8.0	57	7.1 - 8.0
09	8.1 - 9.0	58	8.1 - 9.0
10	9.1 - 10.0	59	9.1 - 10.0
11	10.1 - 15.0	60	10.1 - 15.0
12	15.1 - 20.0	61	15.1 - 20.0
13	20.1 - 25.0	62	20.1 - 25.0
14	25.1 - 30.0	63	25.1 - 30.0
15	30.1 - 35.0	64	30.1 - 35.0
16	35.1 - 40.0	65	35.1 - 40.0
17	40.1 - 45.0	66	40.1 - 45.0
18	45.1 - 50.0	67	45.1 - 50.0
19	50.1 - 75.0	68	50.1 - 75.0
20	75.1 - 100.0	69	75.1 - 100.0
21	100.1 - 125.0	70	100.1 - 125.0
22	125.1 - 150.0	71	125.1 - 150.0
23	150.1 - 175.0	72	150.1 - 175.0
24	175.1 - 200.0	73	175.1 - 200.0
25	200.1 - 300.0	74	200.1 - 300.0
26	300.1 - 400.0	75	300.1 - 400.0
27	400.1 - 500.0	76	400.1 - 500.0
28	500.1 - 600.0	77	500.1 - 600.0
29	600.1 - 700.0	78	600.1 - 700.0
30	700.1 - 800.0	79	700.1 - 800.0
31	800.1 - 900.0	80	800.1 - 900.0
32	900.1 - 1000.0	81	900.1 - 1000.0
33	1000.1 - 1250.0	82	1000.1 - 1250.0
34	1250.1 - 1500.0	83	1250.1 - 1500.0
35	1500.1 - 1750.0	84	1500.1 - 1750.0
36	1750.1 - 2000.0	85	1750.1 - 2000.0
37	2000.1 - 2250.0	86	2000.1 - 2250.0
38	2250.1 - 2500.0	87	2250.1 - 2500.0
39	2500.1 - 3000.0	88	2500.1 - 3000.0
40	3000.1 - 3500.0	89	3000.1 - 3500.0
41	3500.1 - 4000.0	90	3500.1 - 4000.0
42	4000.1 - 4500.0	91	4000.1 - 4500.0
43	4500.1 - 5000.0	92	4500.1 - 5000.0
44	5000.1 - 5500.0	93	5000.1 - 5500.0
45	5500.1 - 6000.0	94	5500.1 - 6000.0
46	6000.1 - 6500.0	95	6000.1 - 6500.0
47	6500.1 - 7000.0	96	6500.1 - 7000.0
48	7000.1 - 7500.0	97	7000.1 - 7500.0
49	7500.1 - 8000.0	98	7500.1 - 8000.0

---

[a] Distance from right bank (facing downstream).

[b] Distance from left bank (facing downstream).

TABLE 4. ALPHABETIC CODES IDENTIFYING SITE (OR PROJECT) AT WHICH SAMPLE WAS COLLECTED (UTILIZED IN COLUMNS 29 AND 30 OF CODING FORM--FIGURE 2)

Site prefix code	Site[a]	Site prefix code	Site[a]
AA	Raccoon Mountain (P)	BW	Chatuge (R)
AB	Bellefonte (N)	BX	Nottely (R)
AC	Browns Ferry (N)	BY	Parksville (R)
AD	Hartsville (N)	BZ	Blue Ridge (R)
AE	Sequoyah (N)	CA	Hales Bar (R)
AF	Watts Bar (N)	CB	Great Falls (R)
AG	Yellow Creek (N)	CC	Barkley (R)
AH	Watts Bar (FF)	CD	Cheatham (R)
AI	Johnsonville (FF)	CE	Old Hickory (R)
AJ	Widows Creek (FF)	CF	Ocoee No. 1 (H)
AK	Shawnee (FF)	CG	Wilbur (H)
AL	Kingston (FF)	CH	Ocoee No. 2 (H)
AM	Colbert (FF)	CI	Nolichucky (H)
AN	John Sevier (FF)	CJ	Great Falls (H)
AO	Gallatin (FF)	CK	Wilson (H)
AP	Allen (FF)	CL	Blue Ridge (H)
AQ	Paradise (FF)	CM	Norris (H)
AR	Bull Run (FF)	CN	Wheeler (H)
AS	Cumberland (FF)	CO	Pickwick (H)
AT	Kentucky (F)	CP	Guntersville (H)
AU	Pickwick (R)	CQ	Chickamauga (H)
AV	Wilson (R)	CR	Hiwassee (H)
AW	Wheeler (R)	CS	Cherokee (H)
AX	Guntersville (R)	CT	Watts Bar (H)
AY	Nickajack (R)	CU	Nottely (H)
AZ	Chickamauga (R)	CV	Chatuge (H)
BA	Watts Bar (R)	CW	Ocoee No. 3 (H)
BB	Fort Loudon (R)	CX	Appalachia (H)
BC	Melton Hill (R)	CY	Douglas (H)
BD	Norris (R)	CZ	Fort Loudon (H)
BE	Douglas (R)	DA	Kentucky (H)
BF	Nolichucky (R)	DB	Fontana (H)
BG	Cherokee (R)	DC	Watauga (H)
BH	Fort Patrick Henry (R)	DD	South Holston (H)
BI	Boone (R)	DE	Boone (H)
BJ	Watauga (R)	DF	Fort Patrick Henry (H)
BK	South Holston (R)	DG	Melton Hill (H)
BL	Tims Ford (R)	DH	Nickajack (H)
BM	Tellico (R)	DI	Tims Ford (H)
BN	Chilhowee (R)	DJ	Clinch R. Breeder
BO	Calderwood (R)	DK	Clinch R. Carbo Plant
BP	Cheoah (R)	DL	Jamestown CH (S)
BQ	Fontana (R)	DM	Jamestown CM (S)
BR	Santetlah (R)	DN	Jamestown CA (S)
BS	Nantahala (R)	DO	Jamestown LY (S)
BT	Thorpe (R)	DP	Jamestown LB (S)
BU	Appalachia (R)	DQ	Jamestown PM (S)
BV	Hiwassee (R)		

[a] Abbreviations: P = pump storage facility; N = nuclear power plants; FF = fossil fuel power plants; R = reservoir; H = hydro plants; and S = strip mine site.

TABLE 5. CODES FOR IDENTIFYING TYPE OF SAMPLE (COMMUNITY, PARAMETER, TEST) UTILIZED IN COLUMN 46 OF CODING FORM--FIGURE 2

Code	Sample type
A	Zoomacrobenthos
B	Periphyton
C	Phytoplankton
D	Zooplankton
E	Macrophyton
F	Productivity, light bottle-dark bottle (oxygen)
G	Productivity, ^{14}C
H	Bioassay, acute
I	Bioassay, chronic

TABLE 6. CODES UTILIZED TO REPORT DATA UNITS (COLUMNS 49 AND 50 ON CODING FORM--FIGURE 2)

Area

A0	Undefined
A1	Sq. micron
A2	Sq. millimeter
A3	Sq. centimeter
A4	Sq. meter
A5	Hectare
A6	Sq. kilometer
A7	Sq. inch
A8	Sq. foot
A9	Sq. yard
AA	Acre
AB	Sq. mile

Chlorophyll-pigments

C0	Undefined
C1	Micrograms active chlorophyll A/sq. centimeter
C2	Micrograms phaeophytin/sq. centimeter
C3	Micrograms chlorophyll A/sq. centimeter
C4	Micrograms chlorophyll B/sq. centimeter
C5	Micrograms chlorophyll C/sq. centimeter
C6	Micrograms beta-carotene/sq. centimeter
C7	Milligrams active chlorophyll A/sq. meter
C8	Milligrams phaeophytin/sq. meter
C9	Milligrams chlorophyll A/sq. meter
CA	Milligrams chlorophyll B/sq. meter
CB	Milligrams chlorophyll C/sq. meter
CC	Milligrams beta-carotene/sq. meter
CD	Milligrams active chlorophyll A/cubic meter
CE	Milligrams phaeophytin/cubic meter
CF	Milligrams chlorophyll A/cubic meter
CG	Milligrams chlorophyll B/cubic meter
CH	Milligrams chlorophyll C/cubic meter
CI	Milligrams beta-carotene/cubic meter
CJ	Milligrams chlorophyll A/liter

TABLE 6 (continued)

Percentages

0	Undefined
X1	Percent abundance, numbers
X2	Percent abundance, biomass
X3	Percent efficiency

Productivity-respiration

P0	Undefined
P1	Milligrams ATP/cubic meter/day
P2	Milligrams ATP/cubic meter/hour
P3	Milligrams ATP/sq. meter/day
P4	Milligrams ATP/sq. meter/hour
P5	Milligrams C/cubic meter/day
P6	Milligrams C/cubic meter/hour
P7	Milligrams C/sq. meter/day
P8	Milligrams C/sq. meter/hour
P9	Milligrams CO_2/cubic meter/day
PA	Milligrams CO_2/cubic meter/hour
PB	Milligrams O_2/cubic meter/day
PC	Milligrams O_2/cubic meter/hour
PD	Milligrams protein/cubic meter/day
PE	Milligrams protein/cubic meter/hour
PF	Milligrams protein/sq. meter/day
PG	Milligrams protein/sq. meter/hour
PH	Grams C/cubic meter/day
PI	Grams C/cubic meter/hour
PJ	Grams C/sq. meter/day
PK	Grams C/sq. meter/hour
PR	Milligrams O_2 uptake/gram fr. wt./day
PS	Milligrams O_2 uptake/gram fr. wt./hour
PT	Milligrams O_2 uptake/gram dry wt./day
PU	Milligrams O_2 uptake/gram dry wt./hour

Radiation-light

R0	Undefined
R1	Foot-candles
R2	Gram calories/centimeter square
R3	Langleys
R4	Lux
R5	Percent of surface illumination

TABLE 6 (continued)

Rate-(l)

10	Undefined
11	Millimeters/second
12	Centimeters/second
13	Meters/second
14	Millimeters/minute
15	Centimeters/minute
16	Meters/minute
17	Millimeters/hour
18	Centimeters/hour
19	Meters/hour
1A	Kilometers/hour
1B	Inches/second
1C	Inches/minute
1D	Inches/hour
1E	Feet/second
1F	Feet/minute
1G	Feet/hour
1H	Miles/hour

Rate-(v)

20	Undefined
21	Cubic feet/second
22	Cubic meters/second

Rate-(wt/a)

30	Undefined
31	Grams/sq. meter/day
32	Grams/sq. meter/hour
33	Milligrams/sq. meter/day
34	Milligrams/sq. meter/hour

Rate-(wt/vol)

40	Undefined
41	Grams/cubic meter/day
42	Grams/cubic meter/hour
43	Micrograms/liter/day
44	Micrograms/liter/hour
45	Milligrams/liter/day
46	Milligrams/liter/hour

TABLE 6 (continued)

Flags and foul-ups

F0	Undefined
F1	Sample lost during collection
F2	Sample lost during analysis
F3	Unable to access sampling station
F4	Unable to recover sample or substrate

Length

L0	Undefined
L1	Micron
L2	Millimeter
L3	Centimeter
L4	Meter
L5	Kilometer
L6	Inch
L7	Foot
L8	Yard
L9	Mile
LA	International nautical mile

Number

N0	Undefined
N1	Number of colonies
N2	Number of eggs
N3	Number of exuviae
N4	Number of hatches
N5	Number of individuals
N6	Number/acre
N7	Number/cubic meter
N8	Number/gram dry wt.
N9	Number/gram ashfree wt.
NA	Number/hectare
NB	Number/liter
NC	Number/milliliter
ND	Number/sq. millimeter
NE	Number/sq. centimeter
NF	Number/sq. foot
NG	Number/sq. meter
NH	Number/sample
NI	Number/cubic meter X 10^3

TABLE 6 (continued)

Ratio (wt/a)

50	Undefined
51	Grams/sq. meter
52	Kilograms/acre
53	Kilograms/hectare
54	Micrograms/sq. centimeter
55	Milligrams/sq. centimeter
56	Milligrams/sq. meter
57	Grams/sq. meter, ashfree
58	Milligrams/sq. centimeter, ashfree
59	Milligrams/sq. meter, ashfree
5A	Grams/sq. meter, ash
5B	Milligrams/sq. centimeter, ash
5C	Milligrams/sq. meter, ash
5D	Kilograms/channel

Ratio (wt/vol)

60	Undefined
61	Grams/cubic meter
62	Grams/liter
63	Kilograms/cubic meter
64	Micrograms/milliliter
65	Milligrams/cubic meter
66	Milligrams/liter
67	Milligrams/milliliter
68	Micrograms/liter

Temperature

D0	Undefined
D1	Degrees Celsius
D2	Degrees Fahrenheit
D3	Degrees Kelvin
D4	Degree-days

Time

T0	Undefined
T1	Day
T2	Hour
T3	Minute
T4	Month
T5	Second
T6	Week
T7	Year

TABLE 6 (continued)

Turbidity

G0	Undefined
G1	Jackson turbidity units (JTU)
G2	Formazin turbidity units (FTU)
G3	Coleman nephlos units (CTU)
G4	Percentage transmittance (%T)

Volume

V0	Undefined
V1	Cubic micron
V2	Cubic millimeter
V3	Milliliter
V4	Liter
V5	Cubic meter
V6	Cubic kilometer
V7	Cubic inch
V8	Ounce
V9	Cubic foot
VA	Cubic yard
VB	Cubic mile

Weight

W0	Undefined
W1	Picogram
W2	Microgram
W3	Milligram
W4	Centigram
W5	Gram
W6	Kilogram
W7	Grain
W8	Ounce
W9	Pound
WA	Ton
WB	Milligram, ashfree
WC	Gram, ashfree
WD	Kilogram, ashfree
WE	Milligram, ash
WF	Gram, ash
WG	Kilogram, ash

TABLE 6 (continued)

Zooplankton

Z0	Undefined
Z1	No. females with 0 eggs per brood chamber
Z2	No. females with 1 egg per brood chamber
Z3	No. females with 2 eggs per brood chamber
Z4	No. females with 3 eggs per brood chamber
Z5	No. females with 4 eggs per brood chamber
Z6	No. females with 5 eggs per brood chamber
Z7	No. females with 6-7 eggs per brood chamber
Z8	No. females with 8-10 eggs per brood chamber
Z9	No. females with 11-15 eggs per brood chamber
ZA	No. females with 16-20 eggs per brood chamber
ZB	No. females with 21-25 eggs per brood chamber
ZC	No. females with 26-30 eggs per brood chamber
ZD	No. females with >30 eggs per brood chamber

TABLE 7. CODES FOR IDENTIFYING BASIC TYPE OF HABITAT FROM WHICH SAMPLE WAS COLLECTED, OR HARDNESS AT WHICH BIOASSAY WAS PERFORMED (USED IN COLUMNS 51 AND 52 OF CODING FORM--FIGURE 2)

Code	Ecologic zone	Code	Substratum	Code	Water Hardness (mg/l as $CaCO_3$)
1	Pool	1	Bedrock	01	10
2	Riffle	2	Boulders	02	20
3	Profundal	3	Rubble (small rocks)	03	30
4	Pelagic	4	Gravel	04	40
5	Littoral	5	Sand	05	50
6	Sublittoral	6	Silt	06	60
7	Eulittoral	7	Clay	07	70
8	Abyssal	8	Marl	08	80
9	Channel	9	Organic detritus (unconsolidated)	09	90
A	Overbank	A	Fibrous peat	10	100
B	Tail water	B	Pulpy peat	11	110
C	Supratidal	C	Organic muck	12	120
D	Intertidal			13	130
E	Subtidal			14	140
F	Splash			15	150
G	Intake canal			16	160
H	Discharge canal			17	170
				18	180
				19	190
				20	200

TABLE 8. CODES FOR RECORDING INSTAR OR SIZE OF ORGANISMS COLLECTED OR USED IN BIOASSAYS (UTILIZED IN COLUMNS 54 AND 55 OF CODING FORM--FIGURE 2)

Code	Interval (mm)	Code	Interval (mm)	Code	Interval (mm)
01	0.001 - 0.10	34	24.1 - 25.0	67	575.1 - 600.0
02	0.11 - 0.20	35	25.1 - 27.0	68	600.1 - 625.0
03	0.21 - 0.30	36	27.1 - 29.0	69	625.1 - 650.0
04	0.31 - 0.40	37	29.1 - 31.0	70	650.1 - 675.0
05	0.41 - 0.50	38	31.1 - 33.0	71	675.1 - 700.0
06	0.51 - 0.60	39	33.1 - 35.0	72	700.1 - 725.0
07	0.61 - 0.70	40	35.1 - 37.0	73	725.1 - 750.0
08	0.71 - 0.80	41	37.1 - 39.0	74	750.1 - 775.0
09	0.81 - 0.90	42	39.1 - 41.0	75	775.1 - 800.0
10	0.91 - 1.00	43	41.1 - 43.0	76	800.1 - 825.0
11	1.1 - 2.0	44	43.1 - 45.0	77	825.1 - 850.0
12	2.1 - 3.0	45	45.1 - 50.0	78	850.1 - 875.0
13	3.1 - 4.0	46	50.1 - 75.0	79	875.1 - 900.0
14	4.1 - 5.0	47	75.1 - 100.0	80	900.1 - 925.0
15	5.1 - 6.0	48	100.1 - 125.0	81	925.1 - 950.0
16	6.1 - 7.0	49	125.1 - 150.0	82	950.1 - 975.0
17	7.1 - 8.0	50	150.1 - 175.0	83	975.1 - 1000.0
18	8.1 - 9.0	51	175.1 - 200.0	84	1000.1 - 1100.0
19	9.1 - 10.0	52	200.1 - 225.0	85	1100.1 - 1200.0
20	10.1 - 11.0	53	225.1 - 250.0	86	1200.1 - 1300.0
21	11.1 - 12.0	54	250.1 - 275.0	87	1300.1 - 1400.0
22	12.1 - 13.0	55	275.1 - 300.0	88	1400.1 - 1500.0
23	13.1 - 14.0	56	300.1 - 325.0	89	1500.1 - 1600.0
24	14.1 - 15.0	57	325.1 - 350.0	90	1600.1 - 1700.0
25	15.1 - 16.0	58	350.1 - 375.0	91	1700.1 - 1800.0
26	16.1 - 17.0	59	375.1 - 400.0	92	1800.1 - 1900.0
27	17.1 - 18.0	60	400.1 - 425.0	93	1900.1 - 2000.0
28	18.1 - 19.0	61	425.1 - 450.0	94	2000.1 - 3000.0
29	19.1 - 20.0	62	450.1 - 475.0	95	3000.1 - 4000.0
30	20.1 - 21.0	63	475.1 - 500.0	96	4000.1 - 5000.0
31	21.1 - 22.0	64	500.1 - 525.0	97	5000.1 - 7500.0
32	22.1 - 23.0	65	525.1 - 550.0	98	7500.1 - 10000.0
33	23.1 - 24.0	66	550.1 - 575.0	99	>10000.1

SECTION 4

METHODS

4.1 DESCRIPTION OF ANALYTICAL TECHNIQUES

After the CIS had been placed in operation, emphasis was initially placed on analysis of selected sets of data representing radically different environmental situations with which the staff had had experience. Three analytical (exploratory) techniques were tested with these data. The first was cluster analysis, which was used (1) to determine the similarity of selected sites along two rivers on the basis of the biotic assemblages found in samples from those sites (Q-mode) and (2) to identify biological communities or associations of species (R-mode).

The second technique was an ordination procedure called nonmetric multidimensional scaling (MDS). This technique was used because it allows "one to examine a scatter diagram displaying a summary of the structure of the data without having to first assume that clusters are present" (Rohlf 1970). Ordination was thus an alternative to cluster analysis and served as a check to help determine if stations or species actually formed distinct groups.

The third technique considered was an index of species diversity, which concurrently examines the number of species and the uniformity or evenness of distribution of individuals among species (Pielou 1969).

4.1.1 Cluster Analysis

Cluster analysis is a multivariate analytical technique used to consider simultaneously all the data contained in a large data matrix. A frequent application of this technique is to search for patterns in data, especially data that do not meet assumptions of rigorous statistical methods.

When cluster analysis is used to analyze data from limnological surveys, the data are tabulated by taxa for each station into a data matrix in which rows are taxa and columns are stations at which samples were collected or, in the case of multiple samples from the same station, the samples themselves. The data can be presented as the number of individuals of each species in each sample, presence or absence of species in samples, or ranked abundances of species. A typical matrix for presence-absence data is shown below, in which 1 stands for presence and 0 stands for absence.

Taxa	Station 1	2	3	4	5
A	1	1	0	0	1
B	1	1	0	0	0
C	1	1	1	1	0
D	1	1	0	0	0
E	1	0	1	1	0
F	0	1	1	1	1

Once a data matrix has been compiled, a similarity or distance matrix is computed that expresses the resemblance between each pair of samples or species in the data matrix. Pairwise comparison between columns (samples) is referred to as Q-mode analysis; comparison between rows (species) is called R-mode analysis. Any of a variety of similarity, correlation, or distance coefficients may be used to quantify the resemblance. One of the simplest is Jaccard's coefficient, S_J (Jaccard 1908):

$$S_J = \frac{a}{a + b + c} ,$$

where, in Q-mode analysis;

a = number of taxa found at both stations,
b = number of taxa found at the first station and not the second,
c = number of taxa found at the second station and not the first.

The similarity matrix that results from Q-mode analysis of the hypothetical data matrix on the previous page is:

Stations	1	2	3	4	5
1	-				
2	4/6	-			
3	2/6	2/6	-		
4	2/6	2/6	3/3	-	
5	1/6	2/5	1/4	1/4	

The next step in the procedure is the clustering itself. One of the most widely used techniques for clustering is the unweighted pair-group method using arithmetic averages (UPGMA) (Sokal and Sneath, 1963). Using this procedure the computer first seeks mutually closest resemblances in the similarity or distance matrix. In our example the closest resemblances, indicated by the highest similarity coefficients, are (1) station 1 with station 2 (4/6 = 0.67) and (2) station 3 with station 4 (3/3 - 1.00). Note that station 5's highest similarity (2/5) is with station 2 but that this is not a mutually highest similarity because station 2 is more similar to station 1 than it is to station 5. After the mutually most similar pairs have been found, the average similarities of these stations with all others in the matrix is found, and the matrix is recomputed. Mutually highest pairs in the new similarity matrix are sought, and the process is repeated until all stations have joined a cluster. Continuing with our example:

Avg. sim. (1-2) with (3-4) = (1 w/ 3 + 1 w/ 4 + 2 w/ 3 + 2 w/ 4)/4

= (2/6 + 2/6 + 2/6 + 2/6)4 = 0.33

Avg. sim. (1-2) with (5) = (1 w/ 5 + 1 w/ 5)/2

= (1/6 + 2/5)/2 = 0.28

Avg. sim. (3-4) with (5) = (3 w/ 5 + 4 w/ 5)/2

(1/4 + 1/4)/2 = 0.25

The resulting recomputed similarity matrix, with mutually highest similarity underlined is:

Stations	1-2	3-4	5
1-2	-		
3-4	0.33	-	
5	0.28	0.25	-

Continuing:

Avg. sim. (1-2-3-4) with (5) = [(1-2 w/ 5 + (3-4) w/ 5]/2

= (0.28 + 0.25)/2 = 0.27

The scale at the top of the figure indicates the level of average similarity between stations and clusters of stations. For example, station 1 clusters with station 2 at a similarity of 0.67; they in turn cluster with stations 3 and 4 at an average similarity of 0.33, and the four stations join station 5 at an average similarity of 0.27.

The last step in cluster analysis is to compute a coefficient of cophenetic correlation (Sokal and Rohlf, 1962). This procedure, described by Roback et al. (1969),

> "is necessary because the clustering method involves averaging of similarities in order to express the multidimensional Jaccard coefficient matrix as a 2-dimensional, hierarchical relationship. Sokal and Rohlf (1962) have developed a method of making this comparison in which similarity values from the dendrogram are expressed as a matrix of cophenetic values. A correlation coefficient is computed between this matrix and the original matrix of coefficients of association. This correlation coefficient, called the cophenetic correlation, is a measure of the amount of distortion introduced by the clustering method. The unweighted pair-group method commonly yields a higher cophenetic correlation than other clustering methods."

The principal difficulty in using cluster analysis, particularly in the study of lotic and semilotic systems, is that most agglomerative clustering algorithms construct hyperspheroidal clusters. In some instances, opposite sides of a stream may be less similar to each other than either side is to upstream or downstream areas. In such situations, longitudinal environmental gradients may exist rather than true clusters, and hyperspheres may be poor descriptors. In essence, the use of cluster analysis superimposes hyperspheroidal clusters onto the natural system, whether or not such clusters express the true similarities in the natural system (Kaesler, 1970).

If hyperspheroidal clusters do not exist, the use of cluster analysis will result in distortion of the similarities among samples. Fortunately, the overall amount of distortion introduced by clustering can be measured by the coefficient of cophenetic correlation, r_{cc} (Sokal and Rohlf, 1962; Farris, 1969; Kaesler, 1970). Use of the coefficient provides a means of quantifying the uncertainties about applying cluster analysis to a particular set of data. Values of r_{cc} larger than 0.8 usually indicate that serious distortion has not been introduced. The larger the matrix of similarity coefficients, the greater the chance of introducing distortion of similarities that may go undetected.

4.1.2 Ordination by Nonmetric Multidimensional Scaling

Ordination is an exploratory analytical technique often used when dealing with large sets of data. In aquatic ecology, ordination involves plotting either the samples or the taxa found in the samples in a two- or three-dimensional scatter diagram. The choice of axes on which the points are plotted depends on the method of ordination selected. In the most successful ordinations, either the samples are plotted in a space where the axes are taxa, or the taxa are plotted in a space defined by the samples. The primary advantage of ordination over cluster analysis is that it "allows one to examine a scatter diagram displaying a summary of the structure of the data without having to first assume that clusters are present" (Rohlf, 1970). In this sense, ordination is a powerful alternative to cluster analysis.

Several techniques of ordination are available, including principal component ordination, principal coordinate ordination, the Bray-Curtis polar ordination technique (Whittaker, 1975), and nonmetric multidimensional scaling (MDS) (Kruskal, 1964a & 1964b). Only MDS was dealt with in this project because it seems to be ideally suited to the kinds of data obtained from biological surveys. Specifically, MDS is computationally robust when data are missing; it can accommodate quantitative, ranked, or presence-absence data; and it can be used with any measure of correlation, similarity, or distance.

In contrast, principal component ordination is usually inappropriate with missing data and operates only on matrices of correlation coefficients. Principal coordinate ordination is not hampered by missing data and operates on a distance matrix, but it gives results that are proportional to principal component ordination when no data are missing. Thus, principal coordinate ordination is likely to be redundant when used with principal component ordination.

An important advantage of MDS, according to Sneath and Sokal (1973), is that "it seems better than principal component analysis in giving balance between the large intercluster distances and the fine differences between members of a given cluster." That is, MDS is a reasonable compromise between cluster analysis, which recognizes fine, intracluster distances, and principal component ordination, which gives a better indication of large intercluster distances than of fine intracluster distances.

The following discussion of MDS is based on the work of Green and Carmone (1970) and Rohlf (1972). Principal component analysis, principal coordinate analysis, and the ordinations developed from them operate in such a way as to find dimensions or components that explain the greatest amount of variance or scatter in the data. The first principal component explains the most variance. The second principal component explains as much of the remaining variance as possible, and it is perpendicular to and uncorrelated with the first component. If any variance remains, additional components are computed and arranged in the same manner described for the first two. In practice, the investigator usually settles for three principal components because three is the greatest number that can be expressed graphically although Green (1979) has urged caution because much of the variation in a data set "may be irrelevant to the purposes of the study."

Table 9 shows data on proportional faunal composition of twelve samples from a hypothetical ecosystem containing ten species. Table 10 shows a distance matrix computed from the data in Table 9 after they were transformed by means of an arcsine transformation. The distance matrix is in the Q-mode and thus shows similarities among samples based on their faunal composition.

Figure 3 shows an ordination that resulted from one-dimensional nonmetric MDS, in which the stations are arrayed along a line. The stress equals 0.230, which indicates a fit that is less than fair.

Figures 4 and 5 show the results of nonmetric MDS in two dimensions. The figures show ordinations, respectively, after 1 and 20 iterations. The stress

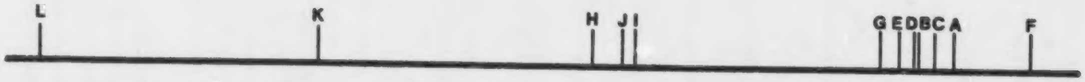


Figure 3. One-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, nine iterations (stress = 0.230).

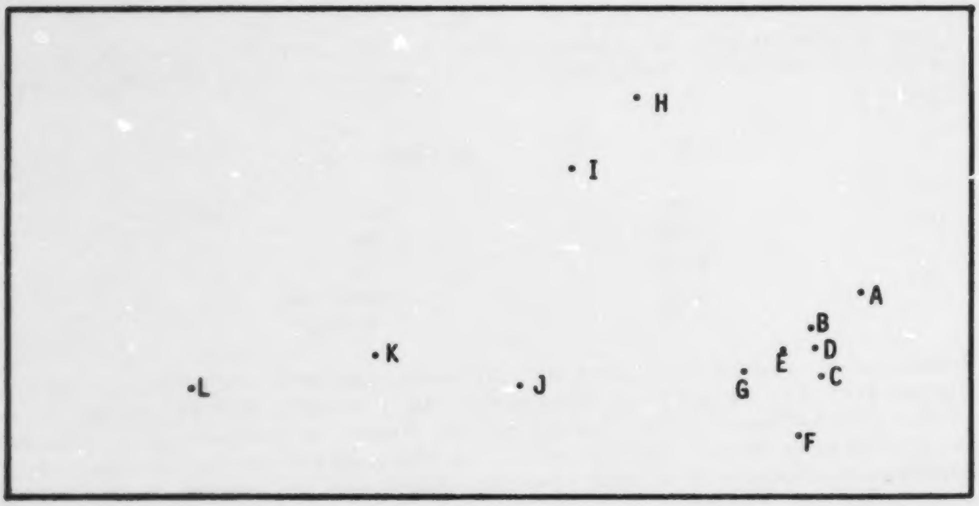


Figure 4. Two-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, one iteration (stress = 0.097).

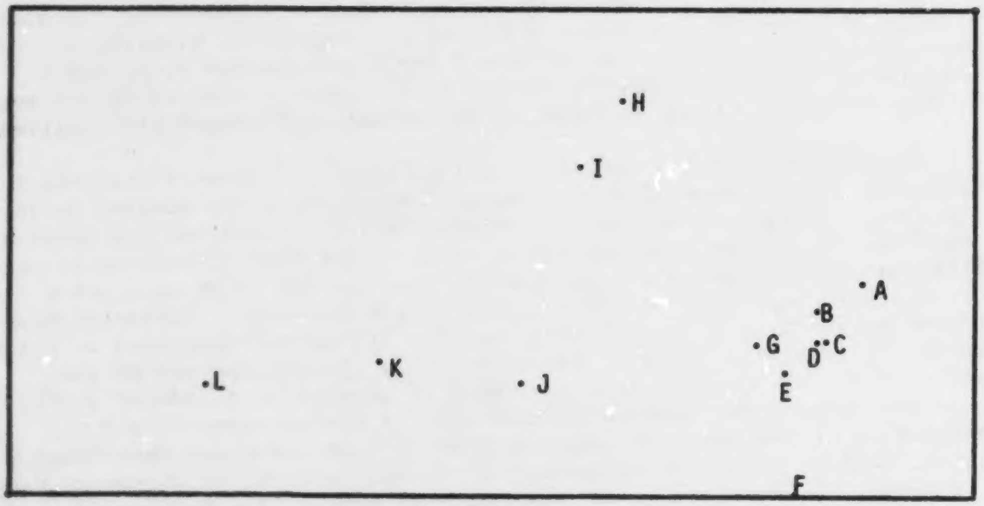


Figure 5. Two-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidensional scaling, 20 iterations (stress = 0.051).

of these ordinations is, respectively, 0.097 and 0.051. In the write-up for the NT-SYS programs, Rohlf evaluated stress as follows (modified from Kruskal, 1964a):

STRESS	GOODNESS OF FIT
0.40	Poor
0.20	Fair
0.10	Good
0.05	Excellent
0.00	"Perfect"

Because our hypothetical data set is small and well structured, the goodness-of-fit is good to excellent. As a result, little change in stress or in the ordination itself is seen as the number of iterations increases. Slight changes in the configuration of the ordination can be observed, however, particularly within cluster ABCDEG and as station F gradually moves away from that cluster.

The ordination in Figure 6 is nearly perfect. It is a three-dimensional ordination in which the vertical axis has been greatly exaggerated. Note that samples A, B, C, D, E, and G are grouped together and that sample F is close on the two horizontal axes. The wide separation of sample F from samples A, B, C, D, E, and G on the vertical axis results partly from the exaggerated vertical scale and partly from its low similarity to the other samples (Table 13). The other five samples (H, I, J, K, and L) are arrayed linearly more than they are grouped into distinct clusters.

For comparison the results of a Q-mode cluster analysis of the same data set (Tables 9 and 10) are shown in Figure 7. Samples A, B, C, D, E, and G form a compact cluster, as do samples H and I and samples J, K, and L. Sample F is similar to cluster ABCDEG, but it is not closely similar to all members of that cluster. (Refer to Table 10 for actual, unaveraged similarities.)

The use of principal component and principal coordinate analysis in an ecological context generally consists of representing the samples collected in a space of reduced dimensionality (Q-mode) where the axes of the space are composites of species that explain as much of the sample variance as possible. Alternatively, one may plot the species found in the study in a space where the axes are composites of samples (R-mode application). Interpretation of an ordination based on either of these methods (principal component or principal coordinate analysis) involves two steps: (1) the explanation of the configuration of points (the arrangement of samples in a reduced species-space or of the species in a reduced reification of station-space), and (2) reification of the axes and relating them to some important environmental factor, such as degree of environmental stress, season, or stream gradient. In general, the development of computational methods for multivariate analysis is far ahead of the development of means of interpreting the results.

Nonmetric MDS takes an entirely different approach to ordination. Similarities or distances between samples are treated on the ordinal scale; that is, they are ordered from smallest to largest. As a result, a configuration is found in which "rank order of (ratio scaled) distances best produces the original input ranks. One tries to do this in the lowest

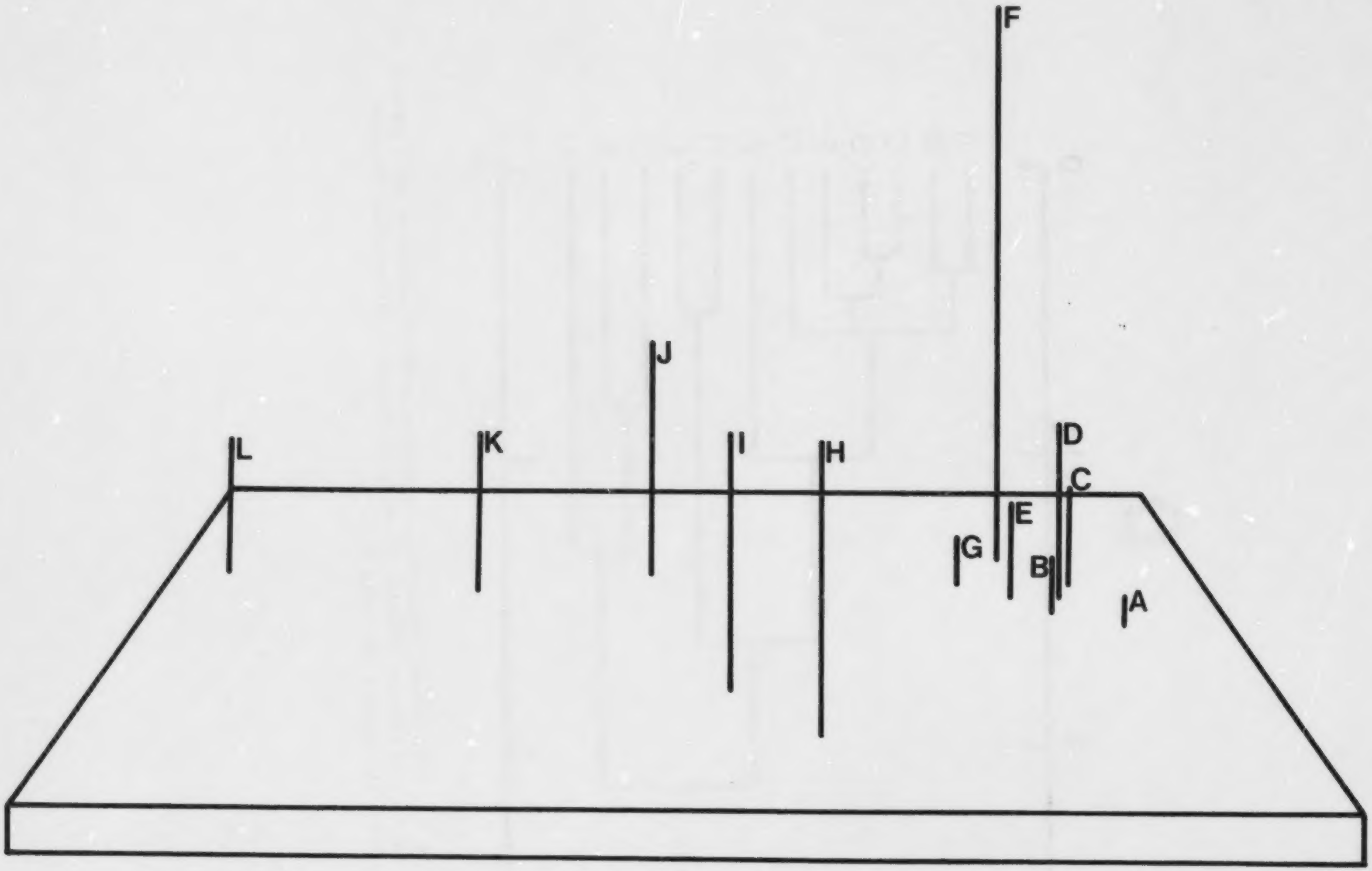


Figure 6. Three-dimensional Q-mode ordination of hypothetical samples in Table 9 computed with nonmetric multidimensional scaling, 43 iterations (stress = 0.001).

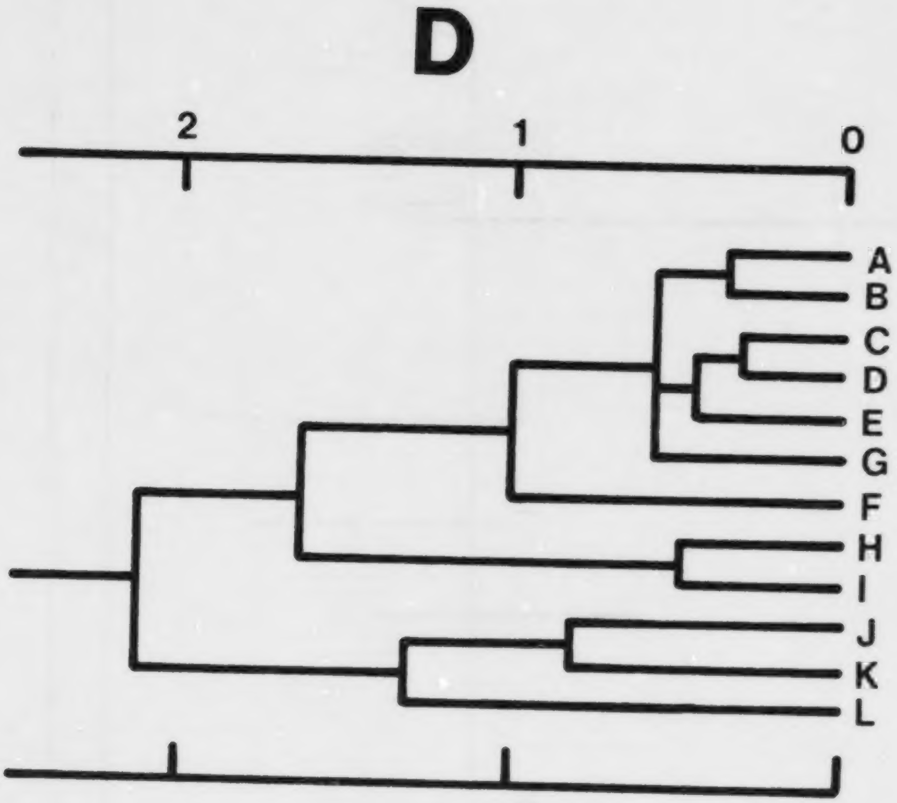


Figure 7. Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients (Table 10) showing faunal similarities between hypothetical samples in Table 9.

dimensionality that produces a 'close enough' ordinal fit" (Green and Carmone, 1970). The important point is that MDS operates on ranked similarities and distances rather than on actual similarities. Thus, MDS is perfectly applicable to matrices of similarity and distance that are computed from ranked data or even presence-absence data.

For a specified number of dimensions, chosen in advance by the investigator, the computer programs "try to find a configuration of points whose interpoint distances are monotone--that is, have the same (or possibly the inverse) ranks as the input data" (Green and Carmone, 1970). The coordinates of this new configuration are the values used in the ordination. In practice, perfect configurations are unusual. The measure of departure from monotonicity is called stress. The higher the stress, the less nearly perfect the degree of monotonicity.

4.1.3 Indices of Species Diversity

Margalef (1956) proposed the use of indices derived in information theory for analysis of multispecies communities. Such use is appropriate where diversity is equated with the uncertainty that exists as to what species will be found when a single organism is selected at random from the community. The greater the number of species present in a community and the more even their distribution, the greater the degree of uncertainty and, thus, the larger the associated species diversity. Information content is a measure of uncertainty and is thus a reasonable measure of diversity as well (Pielou, 1977; Kaesler and Herricks, 1977). The three most commonly used indices of species diversity from information theory are Shannon's index (H') (Shannon and Weaver, 1949), the approximate index (H") (also called $\bar{d}$ by Wilhm and Dorris, 1968), and Brillouin's (1962) index (H).

$$H' = -\sum_{i=1}^{s} p_i \log_e p_i$$

$$H'' = -\sum_{i=1}^{s} (N_i/N) \log_e (N_i/N)$$

$$H = \frac{1}{N} \log_e \frac{N!}{N_1!\, N_2! \, \ldots \, N_s!}$$

where p_i is the probability of selecting at random a member of the ith species, N_i is the number of individuals of the ith species in a collection, N is the total number of individuals of all species in a collection, and s is the number of species. Logarithms to the base e are recommended, although some authors have used the base 2 or the base 10 (Wilhm and Dorris, 1968). It is necessary to specify which logarithmic base is used because the value of diversity is quite base-dependent.

Pielou (1966, 1969, 1975, 1977) and Kaesler and Herricks (1977), in the context of applied aquatic biology, have stressed that Brillouin's index H is the appropriate one for use with fully censused collections of organisms, especially from biological surveys. H' cannot be used because our only knowledge of the p_i's must come from samples. H", which uses data from samples, is a biased estimator of H, always giving too high a value. H, on the other hand, gives the actual species diversity of a fully censused collection of organisms.

In the past, Brillouin's H has been difficult to compute because the factorials involved usually become astronomically large, and even their logarithms are difficult to handle. The ready availability of high-speed digital computers, however, obviates any perceived need to use the approximate diversity H".

In addition to the traditional use of Brillouin's equation for species diversity, diversity was also partitioned into components of diversity using two proposed classification schemes contributed at several levels of three hierarchical classifications, the taxonomic hierarchy, a trophic-functional hierarchy, and a head-body-respiratory functional morphological hierarchy (Tables 11, 12, and 13).

4.2 DESCRIPTION OF DATA SETS

To evaluate and delineate the variations and the usefulness of the three described analytical techniques, three representative sets of zoomacrobenthic data were selected for testing. These are referred to as (1) the Clinch River data set (acute stress); (2) the Cumberland River data set--1973 (chronic stress, high flow); and (3) the Cumberland River data set--1975 (chronic stress, low flow). Zoomacrobenthic data sets alone were utilized in part because Kaesler and Cairns (1972) noted a great deal of redundancy among information derived from the different groups of organisms that are commonly studied as a part of biological surveys. They concluded that the distribution of aquatic insects often is representative of the total biota in lotic environments.

4.2.1 Clinch River Data Set

The Clinch River data set includes information on the fauna collected immediately before and after an accidental release of concentrated sulfuric acid. The spill resulted in an acute stress that killed an estimated 5300 fish. A cursory examination by the Virginia State Water Control Board indicated that stream damage was confined to a 22-km section of the stream, starting 1.5 km downstream from the power plant and extending to St. Paul, Virginia, a distance of 22 km (Soukup, 1970). This conclusion was later substantiated by Crossman et al. (1973, 1974).

Six similar riffle-pool habitats were sampled: one upstream control station, four stations in the affected area, and one station downstream from St. Paul, Virginia (Figure 8). Station 4, 2.5 km upstream from the site of the spill, served as a control station. Stations 7, 8, 9, and 10 were located downstream from the power plant at 3- to 10-km intervals within the affected area. Station 11 was located 11 km downstream from station 10 and was used to substantiate whether the effects of the spill were restricted to the section of the river from the power plant to St. Paul, Virginia.

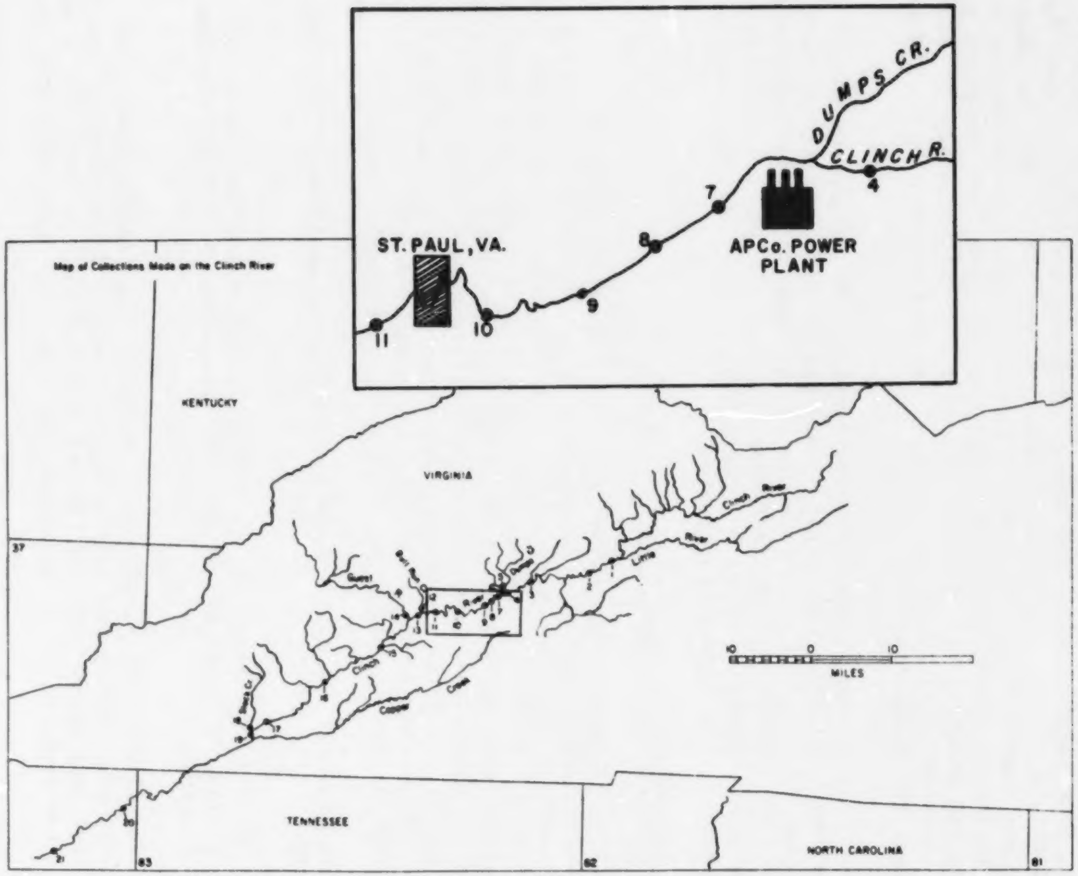

ST. PAUL, VA.
APCo. POWER PLANT
DUMPS CR.
CLINCH R.
11
10
9
8
7
4
Map of Collections Made on the Clinch River
KENTUCKY
VIRGINIA
TENNESSEE
NORTH CAROLINA
Clinch River
Little River
Copper Creek
MILES
10
0
10
37
83
82
81

Figure 8. Map of the Clinch River in Virginia and Tennessee showing the locations of stations sampled during the 1970 zoomacrobenthic survey.

Stations 4, 7, 8, 9, and 10 were sampled 12 to 48 hours before the spill as part of a survey initiated the previous year that also included numerous other upstream and downstream stations (Crossman et al., 1973). Stations 7, 8, 9, 10, and 11 were sampled again immediately after notification of the spill to determine the extent of the damage to the zoomacrobenthic community. Sampling continued at two-week intervals for the next 56 days at stations 7, 8, 9, and 10. Stations 4 and 11, the upstream and downstream control stations, were sampled every four weeks. Physical and chemical data from the power plant and in situ measurements taken in conjunction with biological sampling are summarized in Table 14. Daily measurements of flow from a U.S. Geological Survey (USGS) gauging station 7 km upstream from the power plant are presented in Figure 9. Descriptions of habitats, including information on station location, substratum, width, depth, stream gradient, and riparian vegetation, are summarized in Table 15.

Water quality and stream discharge remained essentially unchanged during the June-August study period (Table 14 and Figure 9). Habitats at each station were also similar, as evidenced by the descriptive data summarized in Table 18. The only major difference was found at station 7. Effluents from the power plant channeled along the right bank at this station, resulting in an alteration of the natural substratum.

4.2.2 Cumberland River Data Sets--1973 and 1975

Two sets of data were utilized from the Cumberland River (a river-reservoir environment). The study area was the headwater region of Old Hickory Reservoir (mean width, 0.9 km), with depths at the different stations varying from 0.9 to 10 m. The major water user within the study area was a steam electric generating plant located on the north bank of the reservoir (Figure 10). Unlike the study area on the Clinch River, this site was affected by the discharges from the power plant's once-through cooling system, which was essentially constant through time. Data were collected from June to October 1973, in January 1975, and from April to September 1975.

Several factors complicated the collection and analysis of the Cumberland River data:

1. The size and shape of the thermal plume varied seasonally and yearly, at times moving upstream over areas previously designated as controls.

2. The flow in the river varied markedly from year to year. For example, during the 1973 study, the median daily average flow did not drop below 5000 cfs for any week, whereas in 1975, there were nine weeks in which median weekly flows were less than 5000 cfs.

3. Some of the stations were located in shallow overbank areas of the reservoir, whereas others were located in the channel (Figure 10).

Rather than sample all benthic habitats, sampling was limited to the predominant silty clay substratum. An a priori assumption was made that the response exhibited by the macrobenthic community inhabiting the silt-clay substratum was representative of the total macrobenthic community.

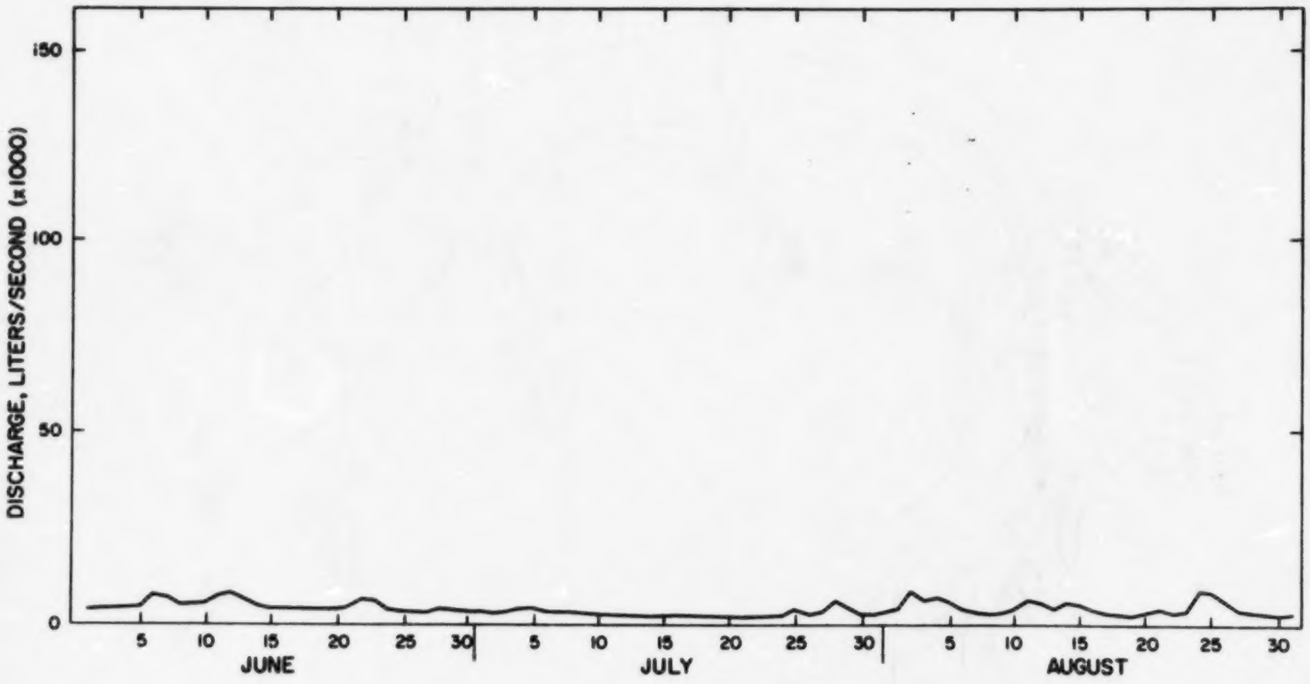


Figure 9. Stream discharge of the Clinch River at the United States Geological Survey gauging station at Cleveland, Virginia, June-August 1970.

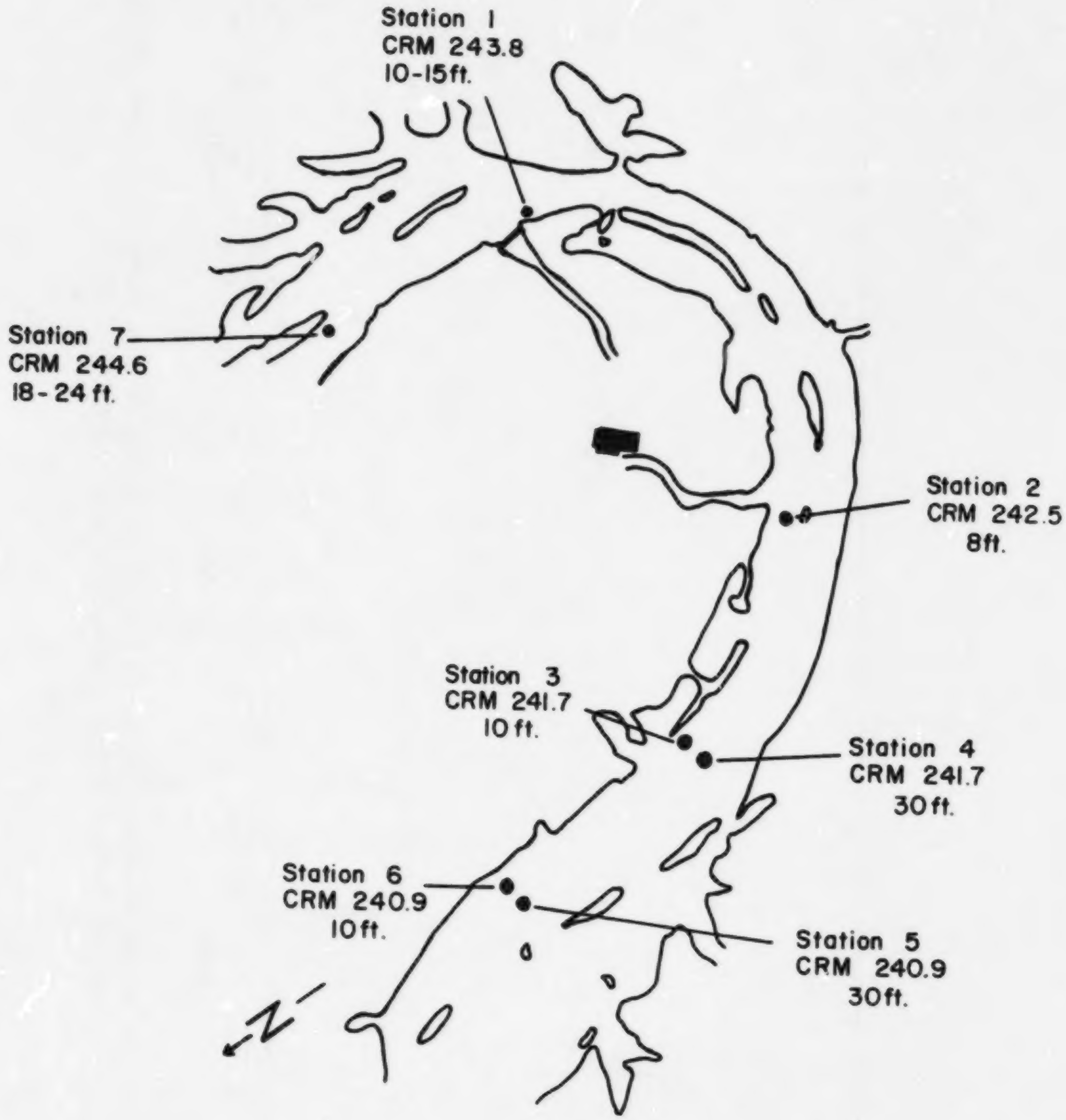

Station 1
CRM 243.8
10-15ft.
Station 7
CRM 244.6
18-24 ft.
Station 2
CRM 242.5
8ft.
Station 3
CRM 241.7
10 ft.
Station 4
CRM 241.7
30 ft.
Station 6
CRM 240.9
10 ft.
Station 5
CRM 240.9
30 ft.
N

Figure 10. Location of stations in the vicinity of a power plant on the Cumberland River.

TABLE 9. HYPOTHETICAL DATA SHOWING PROPORTIONS OF 10 SPECIES AT 12 STATIONS[a]

Species number	Station											
	A	B	C	D	E	F	G	H	I	J	K	L
1	0.110	0.200	0.200	0.200	0.250	0.250	0.300	0.400	0.500	0.600	0.750	0.910
2	0.090	0.100	0.200	0.200	0.250	0.250	0.100	0.050	0.050	0.100	0.050	0.010
3	0.100	0.100	0.100	0.100	0.150	0.250	0.100	0.050	0.050	0.100	0	0.010
4	0.100	0.100	0.100	0.100	0.050	0.250	0.100	0	0	0.100	0.050	0.010
5	0.100	0.100	0.100	0.100	0.050	0	0.100	0	0	0.050	0	0.010
6	0.100	0.100	0.100	0.100	0.050	0	0.100	0	0	0.010	0.050	0.010
7	0.100	0.100	0.100	0.050	0.050	0	0.100	0	0	0.010	0	0.010
8	0.100	0.100	0.100	0.050	0.050	0	0.100	0.050	0.050	0.010	0.050	0.010
9	0.100	0.050	0	0.050	0.050	0	0	0.050	0.050	0.010	0	0.010
10	0.100	0.050	0	0.050	0.050	0	0	0.400	0.300	0.010	0.050	0.010

---

[a] For all subsequent computations, these data were transformed with an arcsine transformation.

TABLE 10. MATRIX OF DISTANCE COEFFICIENTS COMPUTED FROM THE DATA IN TABLE 9 AFTER ARCSINE TRANSFORMATION

	Station											
Station	A	B	C	D	E	F	G	H	I	J	K	L
A	0											
B	0.037	0										
C	0.064	0.039	0									
D	0.055	0.039	0.032	0								
E	0.082	0.064	0.050	0.039	0							
F	0.124	0.108	0.096	0.090	0.082	0						
G	0.076	0.040	0.046	0.056	0.068	0.106	0					
H	0.154	0.149	0.169	0.153	0.147	0.186	0.152	0				
I	0.162	0.147	0.163	0.151	0.141	0.178	0.138	0.049	0			
J	0.181	0.150	0.153	0.149	0.138	0.150	0.119	0.154	0.111	0		
K	0.244	0.215	0.219	0.217	0.206	0.225	0.184	0.182	0.134	0.080	0	
L	0.337	0.308	0.312	0.310	0.297	0.312	0.276	0.265	0.219	0.166	0.098	0

TABLE 11. GENERALIZED TROPHIC, FUNCTIONAL CLASSIFICATION OF ZOOMACROBENTHIC INVERTEBRATES (ADAPTED FROM CUMMINS, 1973)

Level in hierarchy	Name	Subdivision
I	Functional group	Shredders (vascular plant tissues)
		Collectors (detrital materials)
		Grazers (Aufwuchs)
		Predators
		Parasites
II	Feeding mechanism	Chewers and miners
		Filterers (suspension feeders)
		Gatherers (sediment or deposit feeders)
		Scrapers
		Chewers and suckers
		Swallowers and chewers
		Piercers
		Attachers
III	Dependence	Obligate
		Facultative
IV	Food habit	Herbivory
		Detritivory
		Carnivory
		Omnivory

TABLE 12. HIERARCHICAL CLASSIFICATION OF THE TROPHIC-FUNCTIONAL ROLE OF ORGANISMS; INCLUDES ONLY THOSE CATEGORIES THAT OCCURRED IN SAMPLES

Functional group	Feeding mechanism	Dependence	Food habit	Code
1-Shredders	1-Chewers & miners	1-Obligate	1-Herbivory	1111
			2-Detritivory	1112
		2-Facultative	4-Omnivory	1124
2-Collectors	2-Filterers	1-Obligate	2-Detritivory	2212
		2-Facultative	4-Omnivory	2224
	3-Gatherers	1-Obligate	2-Detritivory	2312
		2-Facultative	2-Detritivory	2322
	6-Swallowers & chewers	1-Obligate	2-Detritivory	2612
3-Grazers	4-Scrapers	1-Obligate	1-Herbivory	3411
			2-Detritivory	3412
		2-Facultative	4-Omnivory	3424
	5-Chewers & suckers	1-Obligate	1-Herbivory	3511
		2-Facultative	4-Omnivory	3524
4-Predators	6-Swallowers & chewers	1-Obligate	3-Carnivory	4613
		2-Facultative	4-Omnivory	4624
	7-Piercers	1-Obligate	3-Carnivory	4713

TABLE 13. HIERARCHICAL CLASSIFICATION AND NUMERICAL CODES ASSIGNED FOR ZOOMACROBENTHIC INVERTEBRATES BASED ON FUNCTIONAL MORPHOLOGY: HEAD POSITION, BODY SHAPE, AND RESPIRATORY ORGANS

Level in hierarchy	Name	Subdivision
I	Head position (feeding category)	Hypognathous Prognathous Opisthorphynchous Vestigial or other
II	Body shape (current of stream)	Flattened irregular Flattened oval Flattened elongate Compressed laterally Cylindrical Elongate Short, compact Fusiform Irregular Hemicylindrical or subtriangular
III	Respiratory organs (substratum)	Simple filamentous gills Compound filamentous gills Platelike gills Operculate gills Leaflike gills or organs Respiratory dish Respiratory tube Spiracular gills Caudal chamber Plastron Body integument Tracheal respiration

TABLE 14. RANGE AND MEAN OF PHYSICOCHEMICAL DATA FOR THE CLINCH RIVER, JUNE TO SEPTEMBER 1970

Characteristic	Range	Mean
pH	7.4-8.9	8.2
Temperature, °C	14.4-28.3	21.1
Dissolved oxygen, mg/l	4.6-9.8	7.4
Total hardness, mg/l as $CaCO_3$	127-198	155
Conductivity, µmhos/cm	150-305	246

TABLE 15. DESCRIPTIONS OF HABITATS AT STATIONS 4, 7, 8, 9, 10, and 11 ON THE CLINCH RIVER, 1970[a]

Characteristic	Station					
	4	7	8	9	10	11
River kilometer	934	430	427	423	413	402
Mean depth, m	0.4	0.4	0.4	0.4	0.4	0.4
Mean width, m	60	65	25	60	65	60
Stream gradient, m/km	1.1	1.0	1.7	0.3	0.3	0.9
Composition of substratum[b] (percent)						
Bedrock	5-40	5-60[c]	10-30	10-20	0.10	5-15
Rubble	40-75	20[c]-75	40-70	40-40	70-80	40-70
Gravel	10-20	10-15	10-15	10-15	5-20	15-30
Sand	10-15	10-15	10-15	5-10	0-5	10-15
Maximum rooted vegetation	<——Restricted (<10%)——>			<——Moderate (10-50%)——>		
Dominant streamside vegetation	W	W	C	C	W	C

---

[a] Adapted from a classification developed by Pennak (1971).

[b] Because each station was sampled along the right bank, left bank, and midchannel, the composition is expressed as a range.

[c] The right bank area of station 7 was a solid, calcareous substratum.

[d] Symbols: W = woodland; C = combination of woodland on one side of the stream and brush with herbs and grasses on the other side.

SECTION 5

CLUSTER ANALYSIS

5.1 GENERAL DESCRIPTION

Only in recent years has cluster analysis been applied in aquatic ecology and problems related to water pollution. Cairns and Kaesler (1969, 1971), Cairns et al. (1970), Roback et al. (1969), Kaesler and Cairns (1972), and Kaesler et al. (1971) were the first to use cluster analysis to evaluate the impact of a power plant on a river environment. In a similar study Crossman et al. (1973, 1974) used cluster analysis to describe the response of zoomacrobenthic communities to pH stress and the recovery of a stream from a spill of hazardous materials. Stephenson and Dredge (1976) and Stephenson et al. (1976) also used numerical classification and other methods of analysis to estimate the impact of construction on estuarine fisheries and macrobenthic communities.

The purpose of cluster analysis is to produce a classification that expresses the degree of similarity between the items being classified. The major application of cluster analysis in stream surveys has been to determine the similarities between stations or samples on the basis of their contained biotas. Such analysis is referred to as Q-mode analysis. The use of Q-mode analysis, of course, is not limited to biological applications. It can also be used to compare habitats on the basis of their physical and chemical properties. One such application was Shannon's (1970) use of cluster analysis to group 55 lakes in Florida according to water quality and trophic structure.

R-mode analysis quantifies the similarities between species on the basis of their distribution among the samples studied. Little use has been made of this heuristic technique, although ecologists have long recognized the need to quantify associations of species (Forbes, 1907; Shelford, 1915). Buchanan and Lighthart (1973), however were able to associate water parcels in a eutrophic lake with characteristic assemblages of species. Stephenson et al. (1972) also used R-mode comparison to determine whether Petersen communities that can be characterized by one or two dominant species actually exist in natural systems.

5.2 ANALYTICAL PROCEDURES

5.2.1 Selection of Similarity Coefficients

When considering the use of cluster analysis for summarizing data, the investigator must first select an appropriate similarity coefficient. In this study 26 coefficients were tested to (1) evaluate their usefulness in analyzing environmental data and (2) determine which ones were highly correlated so that redundant expressions could be identified and duplication eliminated. Three kinds of coefficients were tested. Pearson's product-moment correlation coefficient and Sokal's (1961) average taxonomic distance were used with species count data, while a number of similarity coefficients were used with presence-absence data. Some of these coefficients have been analyzed and compared previously (Simpson, 1960; Sokal and Sneath, 1963; and Cheetham and Hazel, 1969) but never in the context of aquatic ecology. Their equations are given in Table 16, where n is the number of species, s_j is the standard deviation of the jth variable, S_jk is the square root of the covariance of the jth and kth variables, X_{ij} is the value of the ith species in jth sample, and a, b, c, and d are terms from a 2 by 2 contingency table (Table 17).

In general, correlation coefficients give a high similarity for samples with species present in the same proportions, whereas distance coefficients give a high similarity (low distance) for samples with species present in the same numbers. The similarity coefficients vary, depending on whether negative matches are included. For a more detailed discussion of these coefficients, the reader should consult Sokal and Sneath (1963).

In addition to the different coefficients, several transformations of quantitative data were evaluated: standardization by rows (species), logarithm of the abundance plus 1, and square root of the abundance plus 0.5. Standardization is the process of transforming each row in a data matrix by subtracting the row mean and dividing by the row standard deviation. When a matrix is standardized, each species has a mean abundance of 0 and a standard deviation of 1. Of course, it is not possible to conceptualize a species with a mean abundance of 0, because it would necessitate negative abundances to offset the positive ones. The justification for standardization is that very abundant species are given proportionally less weight.

The logarithmic transformation of species abundance is a drastic procedure (Table 18). It has the net effect of reducing species abundances to ranks and effectively eliminating the impact of very abundant species.

The square-root transformation is a compromise between the use of untransformed data and the logarithmic transformation (Table 19). Throughout the study, correlation and distance coefficients were computed on untransformed and transformed data. In these computations, the coefficients were labeled as shown in Table 20.

To compare the various coefficients and transformations, matrices of correlation, distance, and similarity were computed for 30 zoomacrobenthic samples. The data were collected at six stations from June to October 1973 in the vicinity of the Cumberland River power plant. All correlation and similarity matrices were then compared by using the coefficient of cophenetic correlation, r_{cc} (Sokal and Rohlf, 1962; Kaesler, 1970; Kaesler and Cairns, 1972; and Kaesler et al., 1974). The r_{cc} is a product-moment correlation coefficient computed between corresponding elements of the matrices. If the matrices have identical or exactly proportional values, they are perfectly correlated and $r_{cc} = 1.0$. If $r_{cc} = -1.0$, the values in the matrices show a perfect negative correlation. After the r_{cc}'s were computed, they were arranged in a correlation matrix (Table 21). This matrix was clustered by using the unweighted pair-group method with arithmetic averages (UPGMA), and a dendrogram was constructed (Figure 11). The dendrogram shows the overall similarities in the matrices and groups those coefficients that produce

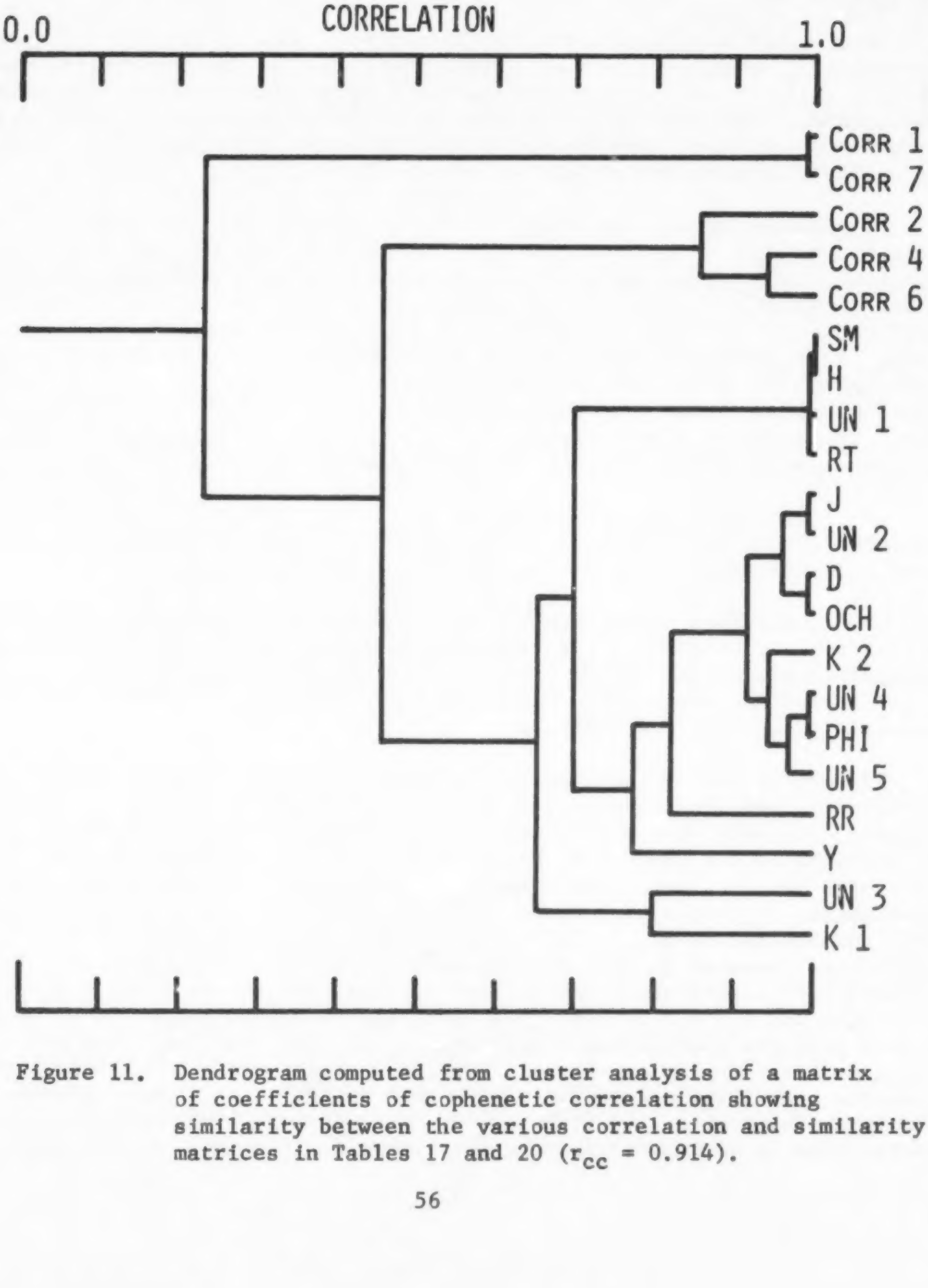

CORRELATION
0.0
1.0
CORR 1
CORR 7
CORR 2
CORR 4
CORR 6
SM
H
UN 1
RT
J
UN 2
D
OCH
K 2
UN 4
PHI
UN 5
RR
Y
UN 3
K 1

Figure 11. Dendrogram computed from cluster analysis of a matrix of coefficients of cophenetic correlation showing similarity between the various correlation and similarity matrices in Tables 17 and 20 ($r_{cc} = 0.914$).

similar matrices. The distance coefficients were not included because they are negatively correlated with similarity matrices with which they are similar because distance is the opposite of similarity.

With 0.75 selected as an arbitrary limit for clustering, five distinct clusters were formed (Figure 11): (1) Corr 1 and 7; (2) Corr 2, 4, and 6; (3) S_{SM}, H, UN1, and RT; (4) S_J, UN2, D, OCH, K2, UN4, RHI, UN5, RR, and Y; and (5) UN3 and K1. One coefficient was selected for further consideration from each of the first four clusters. The low intracorrelations in cluster 5 and the similarity of cluster 5 to clusters 3 and 4 suggested that it could be ignored. The representatives selected from the first four clusters were Corr 7 (square-root transformation, unstandardized data), Corr 6 (square-root transformation, standardized data), S_{SM} (simple matching coefficient), and S_J (Jaccard's coefficient).

The S_{SM} and S_J coefficients were chosen from their respective clusters because of their simplicity and their relatively widespread use in analyzing environmental data. Corr 7 was selected over Corr 1 because of the utility of the square-root transformation. Although Corr 1 and 7 gave similar results, it was possible that some species might be present in very large numbers in other data sets, thus giving misleading results with the product-moment correlation coefficient. Similarly, Corr 6 was selected over Corr 2 and 4 because of the square-root transformation. In this case, however, the transformed data were standardized by rows before the correlation matrix was computed.

A similar rationale was used to select representative distance coefficients (Table 22). Dist 1 and 7 were highly intracorrelated, as were Dist 2, 4, and 6. Dist 7 and 6 were selected as exemplars because of the square-root transformation. The negative correlations between the distance matrices and the correlation and similarity matrices are also presented in Table 22.

5.2.2 Reducing Size of Data Matrices

After the similarity coefficients were selected, data matrices were prepared for the Clinch and Cumberland Rivers data sets. In the case of the Clinch River data set, it was necessary to reduce the size of the data matrix. If all the 1970 Clinch River data had been included (i.e., stations 1 through 21 with right bank, left bank, and midchannel substations), the matrix would have had 248 substations and 123 taxa. This matrix would not only be computationally unmanageable, but the resulting dendrogram would have been difficult to interpret (Figure 12). Two mechanisms were used to limit the size of the data matrices: (1) partitioning the data set into subsets on the basis of the season the stations were sampled and (2) reducing the data set to these samples collected only in immediate vicinity of the spill, i.e., stations 4, 7, 8, 9, 10, and 11.

In order to reduce further the data set to manageable size, the number of taxa needed to be reduced. Relative abundance of species was selected as the criterion upon which to base this reduction. The rationale for using relative abundance was based on Patrick's (1961) discussion on how to determine whether a diatom species was an established resident or a temporary inhabitant of a station. Patrick "considered those species established which were represented

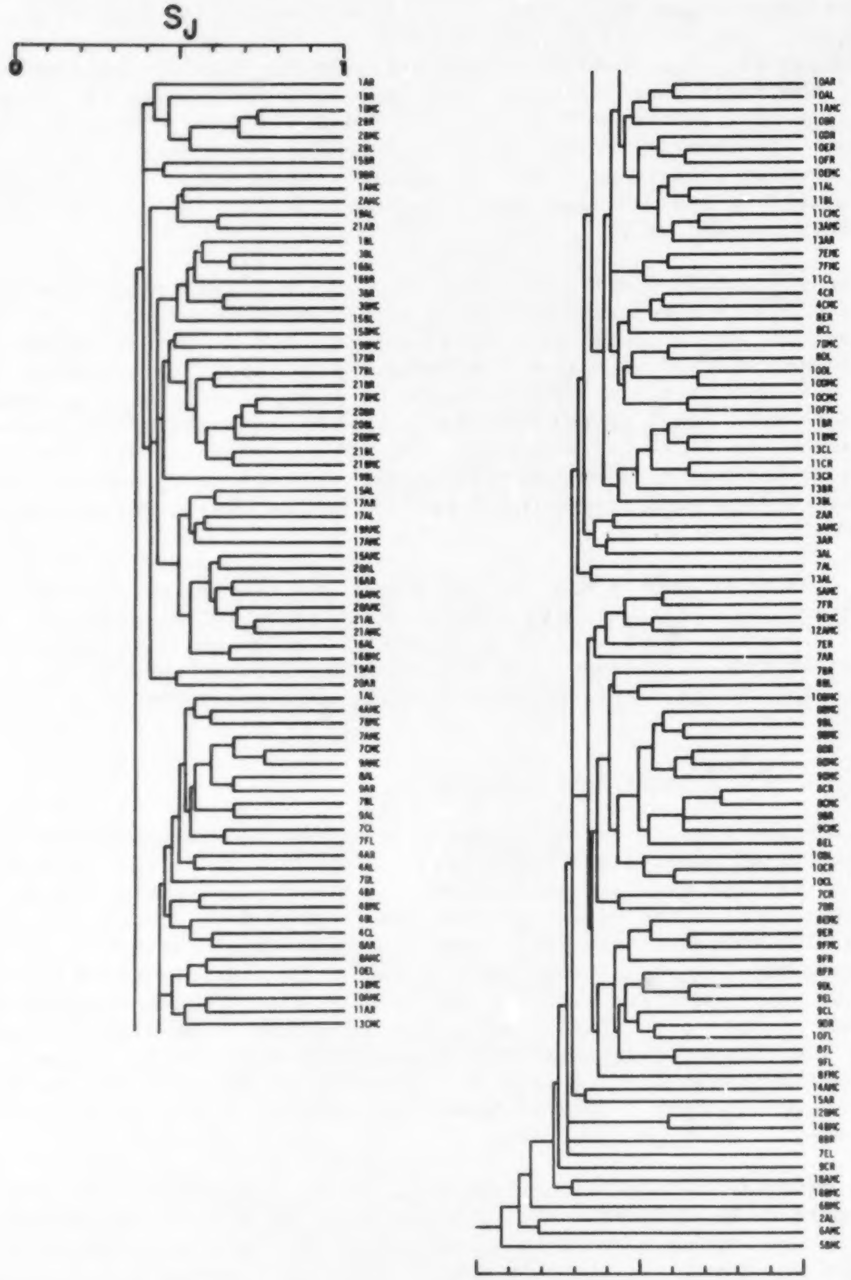


Figure 12. Dendrogram computed from Q-mode cluster analysis of a matrix of Jaccard's coefficients showing faunal similarities between samples collected from the Clinch River in 1970; data include total insect fauna.

by six or more specimens when 8,000 or more diatoms are counted." Since this criterion appeared applicable to any biological community, relative species abundance was selected as the most reasonable means of reducing the number of species.

The formula used for computing relative abundance (RA) was

$$RA = N_i / \sum_{i=1}^{s} N_i ,$$

where N_i = number of individuals for the *i*th species,

$\sum_{i=1}^{s} N_i$ = total number of organisms per station.

To determine the minimum acceptable value of RA, RA's were calculated for each taxon collected at a station unaffected by the spill. Four different levels were considered, 0.10, 0.05, 0.01, 0.005. The 0.10 and 0.05 levels were too exclusive. Emphasis was therefore placed on RA values of 0.01 and 0.005.

An RA of 0.01 reduced the number of taxa from 123 to 29 taxa (Table 23). Representatives of each major taxonomic group normally found in the stream were present. To determine whether major functional groups remained represented, trophic, functional designations were assigned (refer to Tables 11 and 12, section 4.1.3) to each taxon. With the exception of the shredders, every functional group had at least five representatives at the 0.01 level of discrimination (Table 24).

In addition to determining whether major functional groups were present, numerical importance was also considered. Taxa with an RA $\geq$0.01 accounted for 38 to 64 percent of the total number of taxa found at each station (Tables 25 and 26) and 82 to 99 percent of all the organisms found at each site (Tables 27 and 28). While species with an RA <0.01 were important in determining total diversity they were not considered a major component of the stream's macrobenthic community and were, therefore, treated as outliers and deleted from further cluster analyses.

At the RA = 0.005 level of discrimination, the number of taxa increased from 29 to 37 taxa. Since this was an increase of only 8 taxa, the 0.005 level of discrimination was not considered further.

After the taxa to be included had been selected, three data matrices were considered for the Clinch River data set:

1. Those stations affected by the pH stress--stations 7RB[1], 8, 9, and 10 after the spill on June 19, 1970.

2. Those stations unaffected by the spill--stations 4, 7LB&MS[1], and 11 throughout the study and stations 8, 9, and 10 before the spill.

---

[1] Station 7 was divided into 7RB (right bank) and 7LB&MS (left bank and midstream because effluents from the power plant flowed along the right bank.

3. A composite data matrix that considered each station sampled.

The analytical techniques, types of comparison, and similarity coefficients used to analyze each data set and the number of dendrograms and scatter diagrams computed are listed in Table 29.

5.2.3 Evaluation of Distortion

The amount of distortion introduced by the clustering procedure. was measured by the coefficient of cophenetic correlation (r_{cc}). If values were ≥ 0.8, it was assumed that no serious distortion was introduced by the clustering procedure.

5.3 RESULTS

5.3.1 Q-Mode Analysis

5.3.1.1 Clinch River Data Set--Of the six Q-mode dendrograms developed from the Clinch River data set, four had acceptable values of $r_{cc} \geq 0.8$ (Table 30).

5.3.1.1.1 Presence-absence data--The dendrogram from cluster analysis of the matrix of Jaccard's coefficients had the least distortion of any of the Q-mode dendrograms ($r_{cc} = 0.93$). Two clusters were formed at a similarity level of 0.62 (Figure 13). The first cluster contained 22 samples, the second had 6 samples, and eight stations were left unclustered. The first cluster was dominated by samples from stations that were either unaffected by the spill or had recovered from the pH stress (Table 31). The second cluster consisted of samples from stations 8, 9, and 10, the stations impacted by the spill.

A sequence of stream degradation and recovery is indicated by the clusters in Table 31. Station 8, the downstream station nearest the power plant and the most severely impacted site, had only three samples in the first cluster, samples collected before the pH stress and in August, six and eight weeks after the spill. Station 9, the next station downstream, had four samples in the first cluster. They were the same as those for station 8, but also included the late July sample collected four weeks after the spill. At station 10, the farthest downstream station affected by the spill, only the sample collected immediately after the pH stress was missing from the cluster. This indicated that the farther downstream a station was located, the less severe the impact and the faster its recovery.

The six samples in the second cluster are from stations 8, 9, and 10 after the spill. Their clustering indicates that their zoomacrobenthic communities were similar immediately after the spill. The severity of the impact and the time required for recovery depended on the station's proximity to the site of the spill. This cluster also indicates that the Clinch River's zoomacrobenthic community had a high natural resiliency or ability to recover from an acute stress.

The eight unclustered samples were those collected from stations 7RB and station 7LB&MS in August. As noted in the description of the sites, effluents from the power plant channeled along the right bank, resulting in a chronic stress. As a result, the low similarities of samples from station 7RB to other samples were not unexpected. The failure of samples from station 7LB&MS

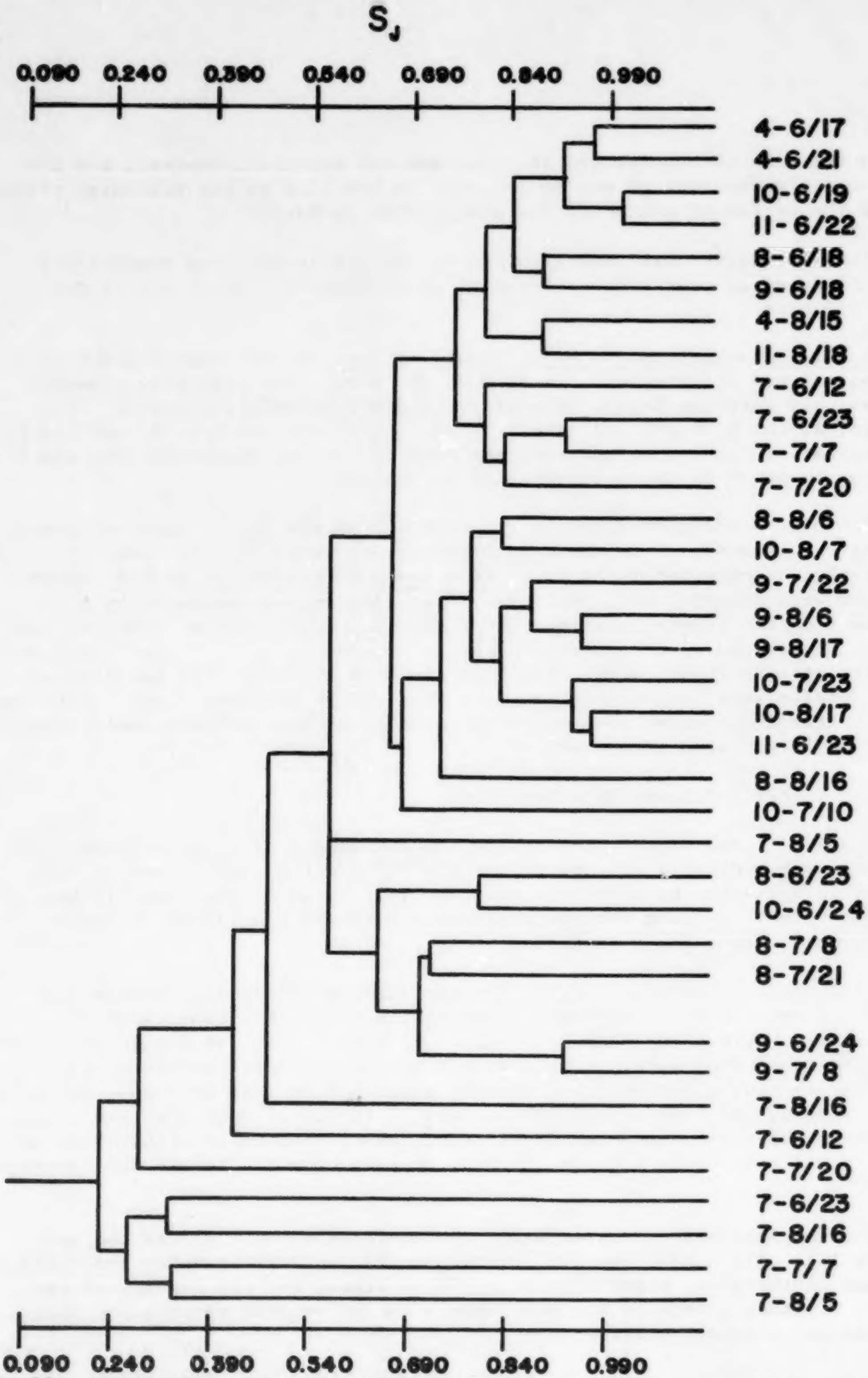


Figure 13. Dendrogram computed from Q-mode cluster analysis of a matrix of Jaccard's coefficients showing faunal similarities between 36 zoomacrobenthic samples collected from stations 4, 7, 8, 9, 10, and 11 in the Clinch River, 1970 (r_{cc} = 0.93).

to cluster with the unaffected stations was not expected, however, and the only possible explanation was an increase in the size of the discharge plume during the period of extremely low stream flow in August.

The dendrogram resulting from use of the simple-matching coefficient (S_{SM}) produced an unacceptable level of distortion ($r_{cc} < 0.8$) and is not discussed here.

5.3.1.1.2 Quantitative data, counts of species--The Corr 6 and Dist 6 dendrograms had acceptable r_{cc}'s of 0.82 and 0.91. The resulting clusters, however, had very low levels of similarity and were uninterpretable. For example, at the 0.06 and 1.2 levels of similarity for the Corr 6 and Dist 6 dendrograms, samples from the upstream control station clustered with the stations impacted by the spill (Tables 32 and 33).

The Dist 7 dendrogram had an r_{cc} of 0.80. Since it is computed from a data matrix transformed by the square-root transformation, the level of similarity was expected to be low. At a level of similarity of 4.8, three clusters were formed (Table 34). The first cluster was dominated by 11 samples from the control stations and from stations 7 through 10 before the spill. The second cluster consisted of 20 samples, from stations 7 through 10, with station 7LB&MS clustering with the most severely impacted stations. These results were contrary to previous findings by Crossman (1973, 1974) and suggest that dendrograms computed from correlation and distance coefficients should be interpreted with caution.

5.3.1.2 Cumberland River Data Set--1973--

5.3.1.2.1 Substrate--Stations on the Cumberland River were established in areas with sediments of similar texture. The intent was to sample only substrates dominated by very fine sand to clay. Cluster analysis of the particle size data using various coefficients showed similarity of the substratum at the various stations.

Cluster analysis of a matrix of correlation coefficients (Figure 14), however, shows changes in sediment composition with time, especially in overbank areas (level of similarity 0.7). A higher level of similarity, 0.82, shows the trend even more strongly with almost all samples collected in September remaining unclustered, whereas samples collected in August tended to join clusters. At both levels of similarity, it was evident that the substratum was more consistent (1) in the channel than in overbank areas and (2) in downstream areas than in upstream areas around the intake and discharge canals.

Cluster analysis of a matrix of distance coefficients showed the same (Figure 15): (1) August samples clustered; (2) September samples generally remained unclustered, especially in overbank areas; and (3) downstream and channel stations generally had more consistent substratum composition than upstream and overbank stations.

5.3.1.2.2 Zoomacrobenthos--Six different similarity coefficients (two presence-absence and four quantitative) were used to examine the total macrobenthic data set for 1973.

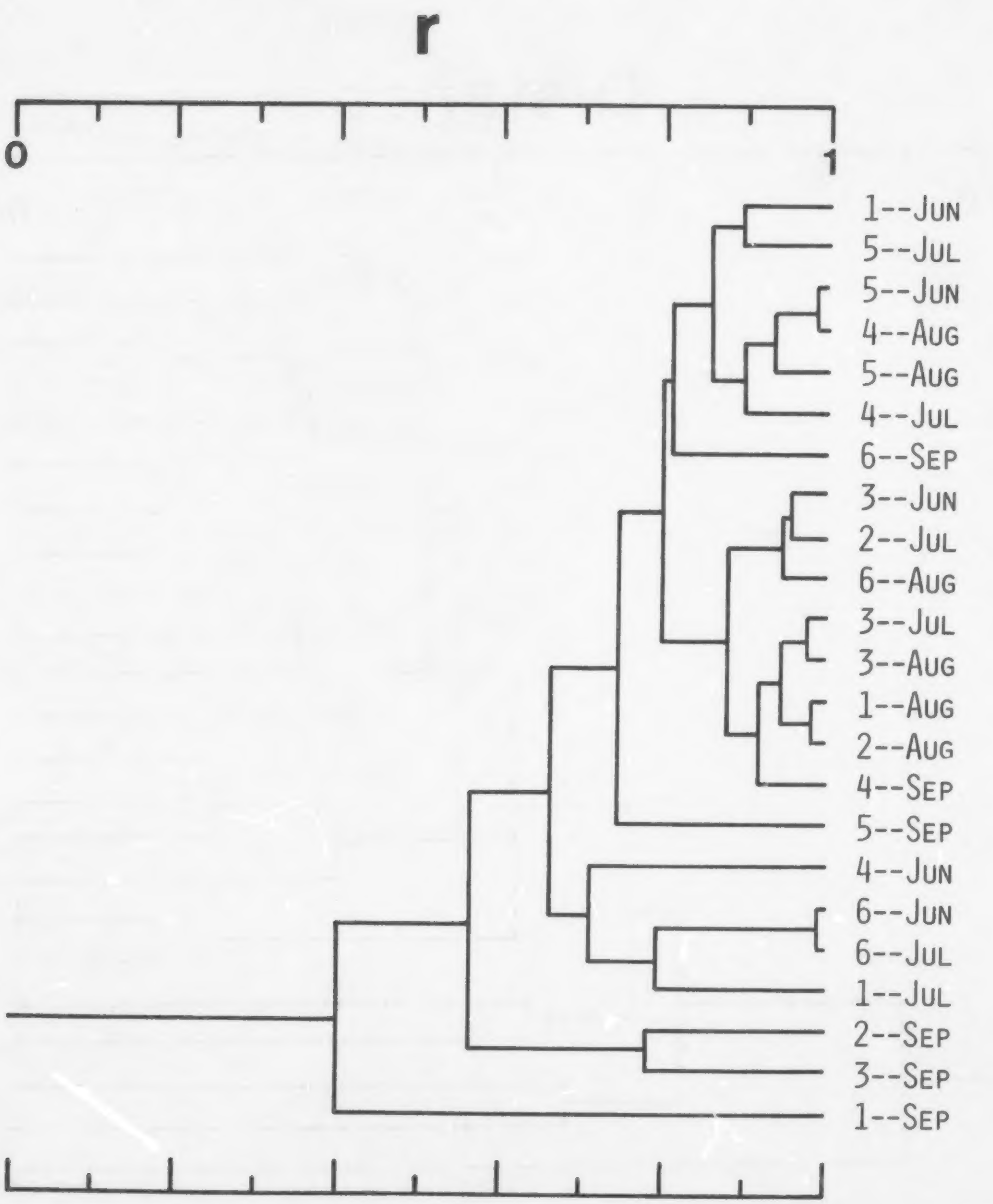


Figure 14. Dendrogram computed from Q-mode cluster analysis of a matrix of correlation coefficients computed from proportions of each phi size in the substrate after arcsine transformation; shows similarity of substrate between samples collected from the Cumberland River in 1973 (r_{cc} = 0.760).

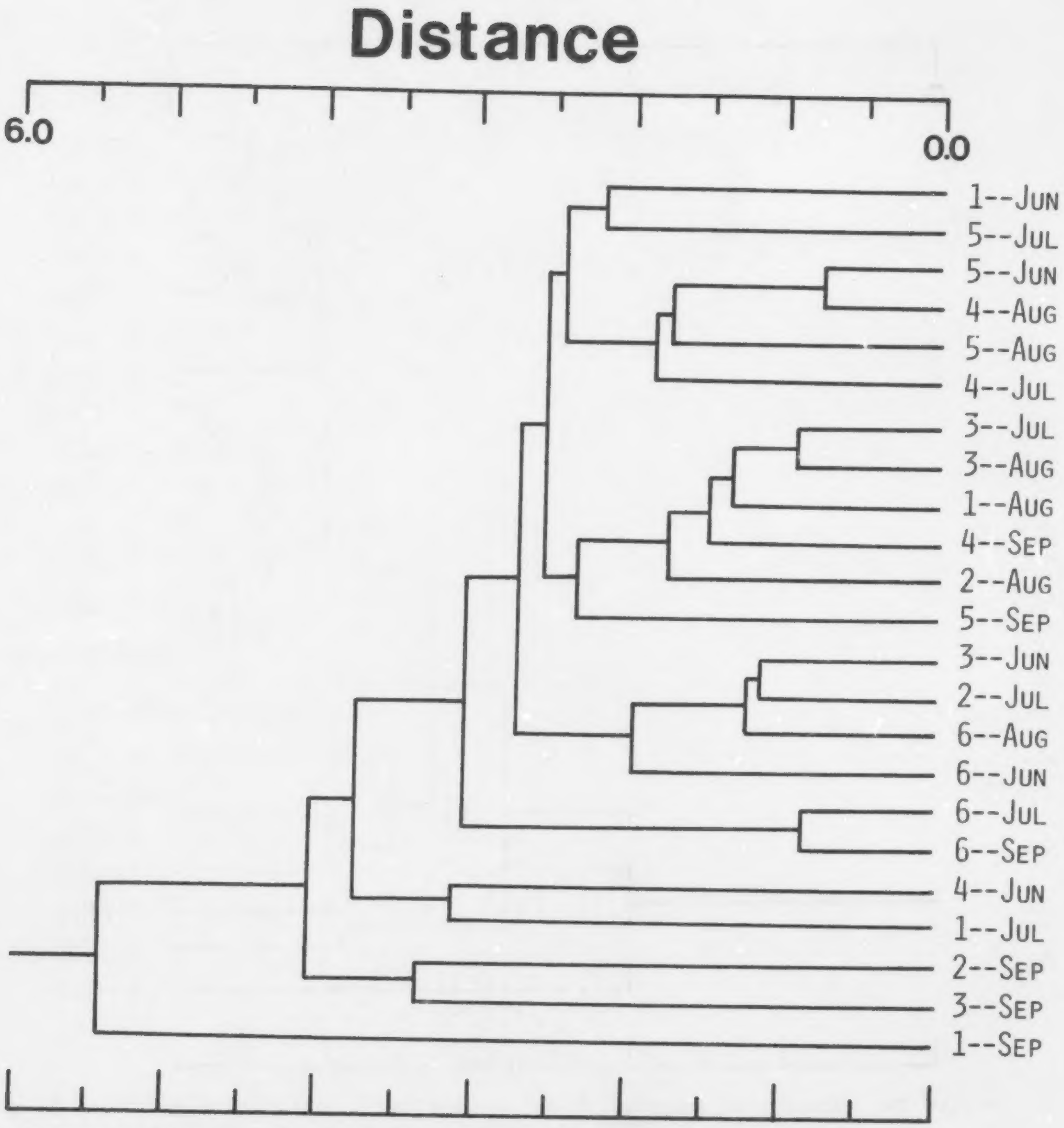


Figure 15. Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from proportions of each phi size in the substrate after arcsine transformation; shows similarity of substrate between samples collected from the Cumberland River in 1973.

5.3.1.2.2.1 Presence-absence data--Jaccard's coefficient (S_J) produced a dendrogram with an acceptably low level of distortion ($r_{cc} = 0.85$). At similarity levels of 0.5 and 0.6, samples from stations 2 and 3 tended to form clusters (Figure 16). These stations were located at the discharge and on the overbank area below the discharge. Moreover, samples collected in September and October tended to form clusters except for samples from stations 2 and 3.

The simple-matching coefficient (S_{SM}) produced higher similarities than Jaccard's coefficient because a number of samples contained rare species (i.e., species that did not occur in more than one or two other samples). The large number of negative matches increased overall similarity. A more thorough discussion of the results obtained with the simple-matching coefficient is not warranted because it included an unacceptable level of distortion ($r_{cc} < 0.8$).

5.3.1.2.2.2 Quantitative data, counts of species--The distance coefficient Dist 7 formed clusters of samples at a high level of similarity and with an acceptably low level of distortion (Figure 17). Two groups of samples appear as distinct clusters at a level of similarity of 2.3: (1) samples collected from station 2 during the summer and from station 1 during the late summer and fall (The temperature at station 1 in late summer and fall was similar to that at station 2 in early summer as a result of warmer weather and possible upstream migration of the thermal plume.); and (2) samples collected at all downstream stations in the month of September. At a higher level of similarity, samples from stations 1 and 2 remained clustered, but two new clusters appeared: (1) samples from stations located in overbank areas only and (2) samples predominantly from the channel stations and entirely from the stations downstream from the plant.

With the correlation coefficient Corr 7, clusters were formed at extremely high levels of similarity (Figure 18). At a level of 0.9, for example, three clusters were formed. The first cluster comprised only samples collected from stations 1 and 2 in June, a similarity that cannot be explained by temperature alone because ΔT between these two stations in June was 5.6°C. The biotic similarity may have resulted from similarity of substrate. The second cluster contained 57 percent of the samples. The only noticeable trend was the complete absence of samples from September, during which time the lowest flows and highest temperatures were recorded. Cluster 3 contained only two samples (3-SEP and 3-OCT), indicating a distinct biota during the autumn at station 3.

5.3.1.2.3 Summary--A thorough discussion of results obtained with Corr 6 and Dist 6 is not warranted because use of these two coefficients produced unacceptable levels of distortion ($r_{cc} < 0.8$). In addition, use of these coefficients produced dendrograms with less structure and lower similarities than those of their counterparts (Corr 7 and Dist 7), indicating that the standardization of data may increase distortion and decrease the structure and similarity level of dendrograms.

A summary of the cluster analyses for the 1973 data set is presented in Table 35. Three of the six coefficients had acceptably low levels of distortion (S_J, Dist 7, and Corr 7). They also produced dendrograms that were readily interpretable, and all three indicated that both station location and the month in which the samples were collected were causes for association.

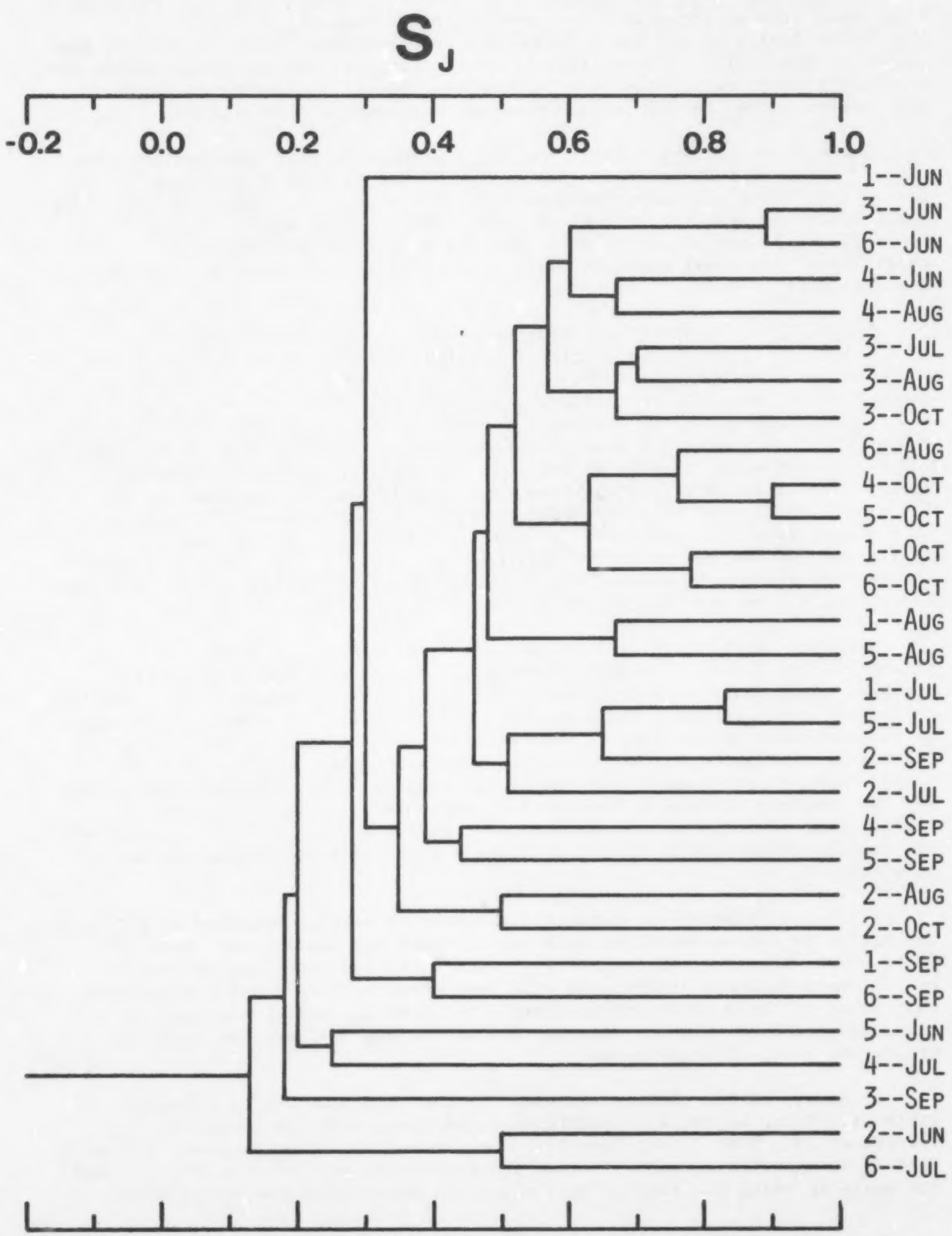


Figure 16. Dendrogram computed from Q-mode cluster analysis of a matrix of Jaccard's coefficients showing faunal similarities between samples collected from the Cumberland River in 1973 ($r_{cc} = 0.852$).

DIST 7

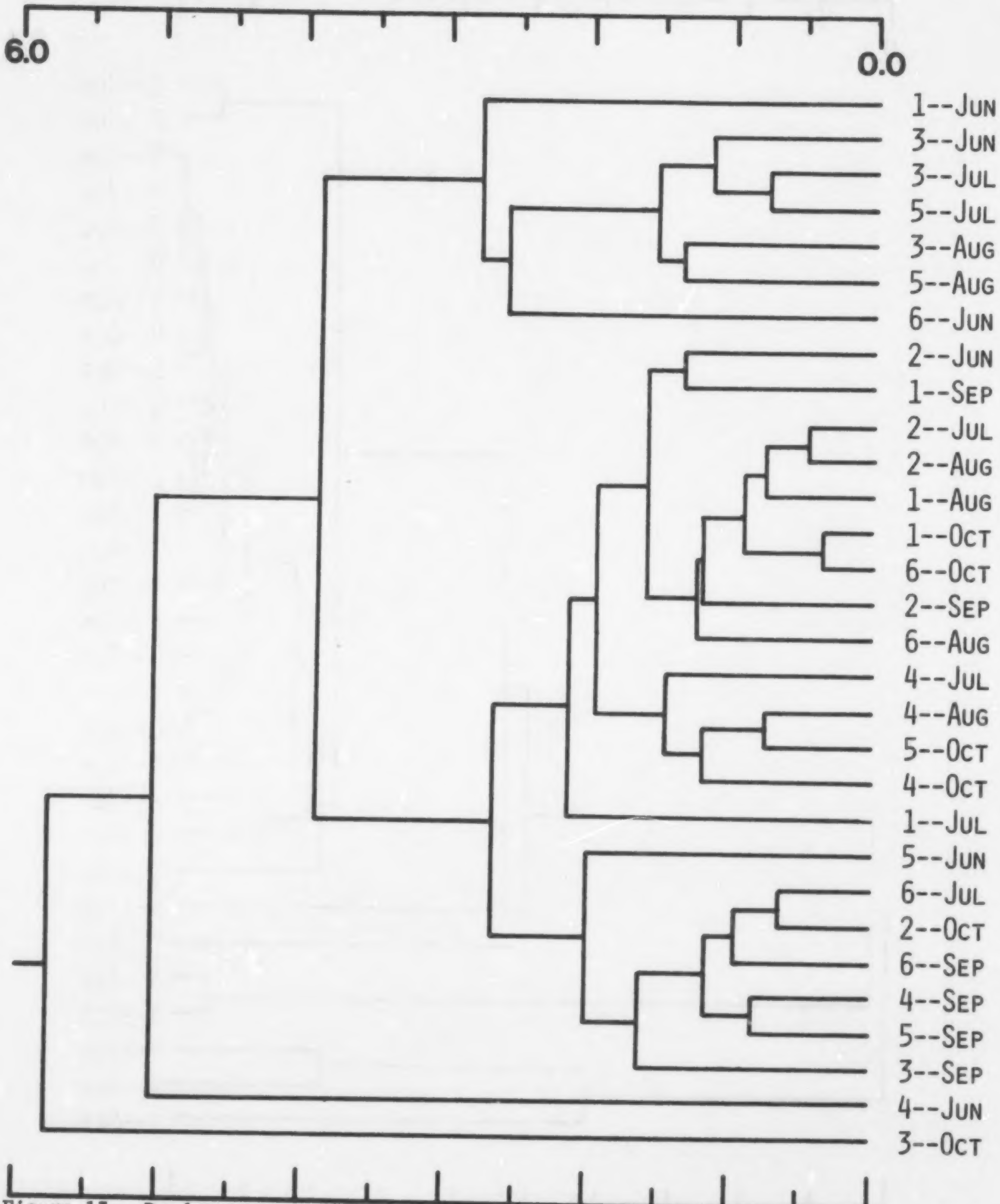


Figure 17. Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from data transformed with the square-root transformation; shows faunal similarities between samples collected from the Cumberland River in 1973 ($r_{cc} = 0.833$).

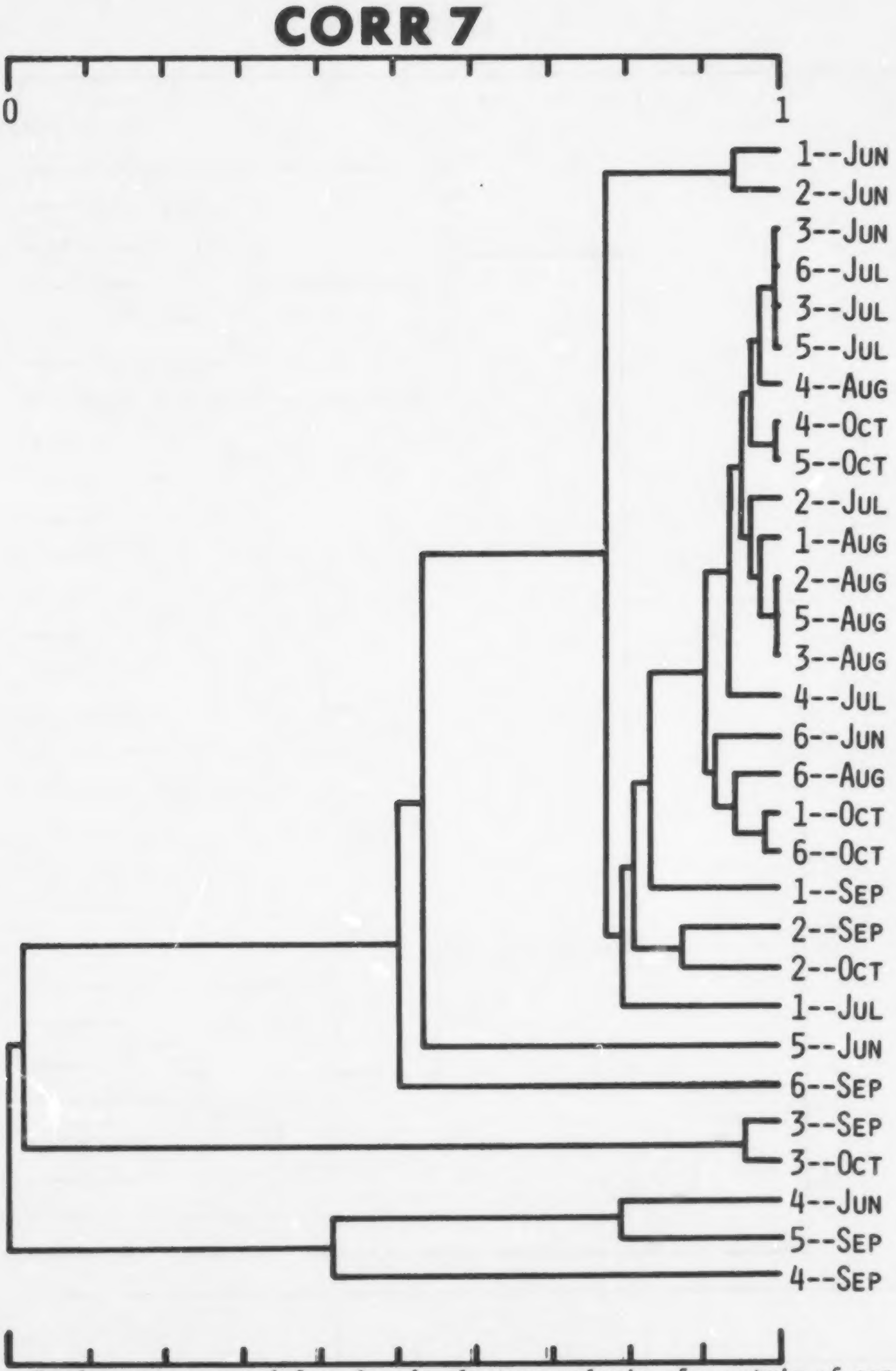


Figure 18. Dendrogram computed from Q-mode cluster analysis of a matrix of correlation coefficients computed from data transformed with the square-root transformation; shows faunal similarities between samples collected from the Cumberland River in 1973 (r_{cc} = 0.979).

Two of the three coefficients whose use produced unacceptable levels of distortion (Corr 6, Dist 6) also produced dendrograms that had low similarities and were not readily interpretable. This result weighs against the use of standardization in the analysis of ecological data.

The high degree of structure and similarity in the dendrogram resulting from use of the simple-matching coefficient was due largely to its inclusion of negative matches, a possible source of bias.

In this case, however, despite the possible bias and the introduction of more distortion than Jaccard's coefficient, the stations associated at the highest levels of similarity were nearly the same with these two presence-absence coefficients (Table 35). The highly similar stations using these presence-absence similarity coefficients were in many cases, however, different from those that were associated at the highest levels by the various quantitative coefficients.

5.3.1.3 Cumberland River Data Set--1975--

5.3.1.3.1 Substratum--Sediment data collected in 1975 from seven stations on the Cumberland River were clustered through use of the Pearson product-moment correlation coefficient. Moderate distortion was introduced through the clustering procedure ($r_{cc} = 0.76$). At a similarity level of 0.40, two large clusters were formed. Cluster 1 contains most of the samples and is dominated by samples from stations 2, 3, and 6, indicating that sediments at the downstream overbank stations were similar. Cluster 2 is dominated by channel stations, suggesting that composition of substratum differed between the overbank and channel sites. At higher levels of similarity, change in sediment composition with time is suggested.

5.3.1.3.2 Zoomacrobenthos--

5.3.1.3.2.1 Presence-absence data--Jaccard's coefficient (S_J) produced a dendrogram that clustered samples at low levels of similarity but with little distortion ($r_{cc} = 0.88$). At a similarity level of 0.68, seven clusters were formed. The first three clusters were dominated by samples collected in May, the fourth cluster was dominated by samples collected in June and July, and the last three clusters were dominated by samples collected in August and September. These groupings suggest that faunal assemblages were influenced more by seasonal factors than by their proximity to the thermal discharge.

This type of distribution is expected when environmental conditions are the same at all stations. Examination of temperature data revealed that temperatures differed very little between the control stations upstream and the stations downstream from the power plant. Temperatures in May were $13 \pm 1.5^{\circ}C$ at all stations, increasing to $26.5 \pm 2^{\circ}C$ at all stations in August.

The simple matching coefficient (S_{SM}) resulted in overall higher similarities, but a great deal more distortion was introduced ($r_{cc} = 0.70$). Because 0.80 had been established as the minimum r_{cc} value for an acceptable dendrogram, interpretation of this dendrogram is limited to the general observation that seasonal factors tended to influence similarities between stations more than location.

5.3.1.3.2.2 Quantitative data, species counts--Cluster analysis with the distance coefficient Dist 6 resulted in a dendrogram with little distortion (r_{cc} = 0.95). At a level of similarity of 0.84, one large cluster and three small clusters were formed, with 11 samples remaining unclustered (Figure 19). The large cluster was dominated by samples collected in May and June and samples from station 3, whereas the unclustered samples were collected predominantly at station 2 or in the month of September.

Only 47 taxa were collected either at station 3 or in May when the diversity was low, whereas 78 taxa were collected at station 2 or in September. Similarly, the mean number of organisms collected per square meter in May was 1,325, while the mean number in September was 2,337. These facts probably account for the separation of these sample groups.

Cluster analysis with the distance coefficient Dist 7 produced a dendrogram with a stair-stepped appearance and little distortion of similarities (r_{cc} = 0.84) (Figure 20). At a similarity level of 1.55, two large clusters and five small clusters were formed. Both large clusters contained samples from (1) at least four of the five months sampled, (2) both overbank and channel stations, and (3) samples both upstream and downstream from the plant. This arrangement seems to indicate that all the samples collected were similar. At higher levels of similarity, however, we were able to detect slight differences with time--a trend that was also noted in cluster analysis of other coefficients.

Cluster analysis of coefficients Corr 6 and Corr 7 resulted in dendrograms that contained an unacceptable amount of distortion and a structure that made interpretations difficult. Neither temporal nor spatial factors were shown to predominate in the formation of biotic associations.

5.3.1.3.4 Summary--Only three of the six coefficients tested, Jaccard's coefficient and the distance coefficients Dist 6 and Dist 7, produced dendrograms with acceptable low levels of distortion ($r_{cc} \geq 0.80$) (Table 36). Samples with highest similarities are nearly the same in all cluster analyses. That is, some samples are closely similar regardless of which coefficient is used. The principal difference among samples with a high similarity was that the two presence-absence coefficients (S_J and S_{SM}) indicated a high similarity between a channel and an overbank sample collected during May, which the quantitative coefficients associated at a much lower level. Similarly, the quantitative coefficients indicated high similarity for one pair of overbank samples in May that the presence-absence coefficients clustered at a much lower level. All three dendrograms with acceptable levels of distortion indicated that the season during which the samples were collected was the primary cause of association. Only the Jaccard coefficient (presence-absence data only), however, failed to indicate that station location was a secondary factor.

For both the 1973 and 1975 Cumberland River data sets, only three of the six coefficients tested produced dendrograms with acceptably low levels of distortion. The Jaccard coefficient and the Dist 7 (quantitative, distance, square root transformation with no standardization) were two of the three that were acceptable for both data sets. For the 1973 data set, the Corr 7 coefficient (quantitative, correlation, square root transformation with no standardization) produced acceptably low distortion, while the Dist 6

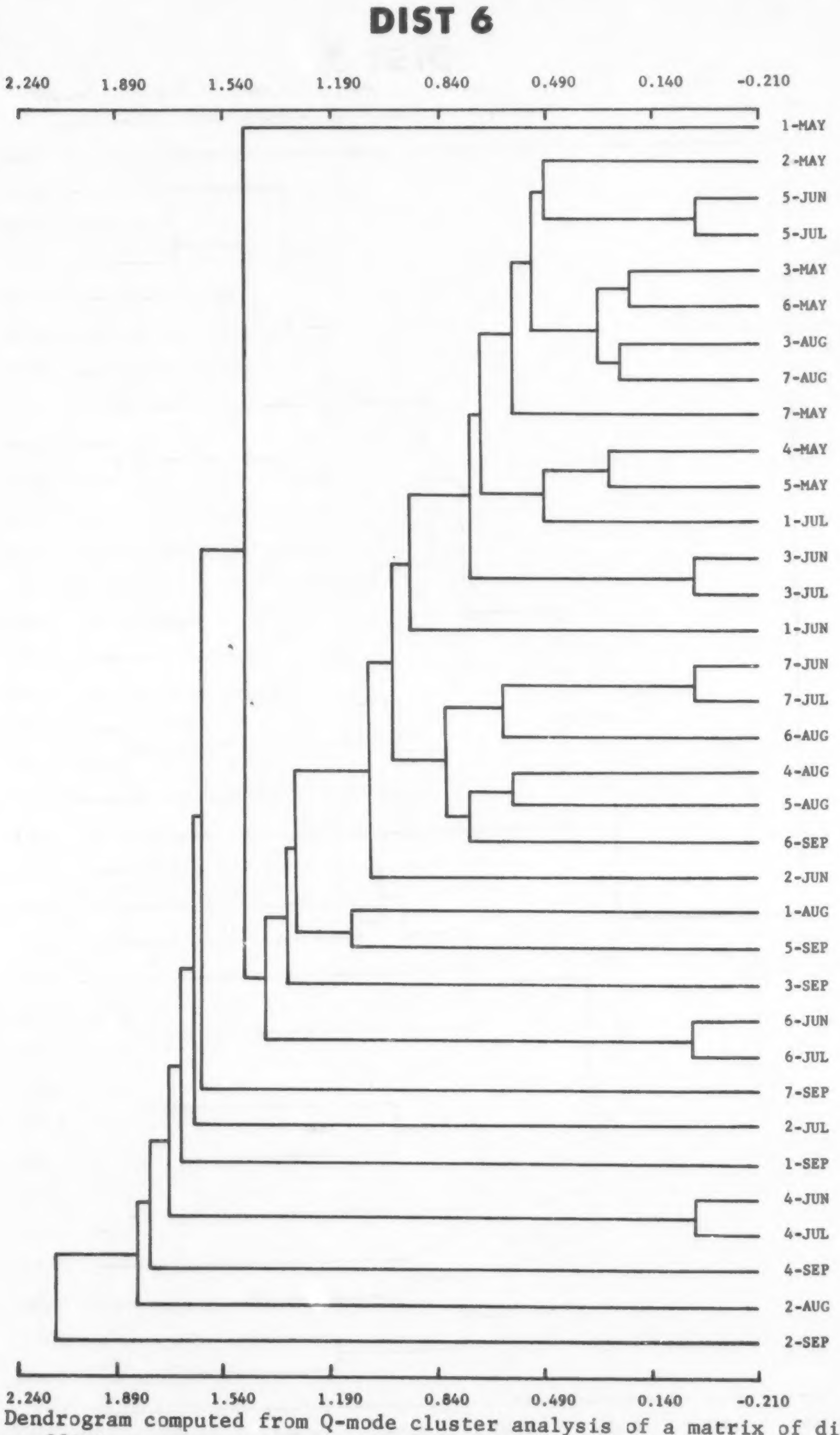


Figure 19. Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation and standardized by rows; shows faunal similarities between samples collected from the Cumberland River in 1975.

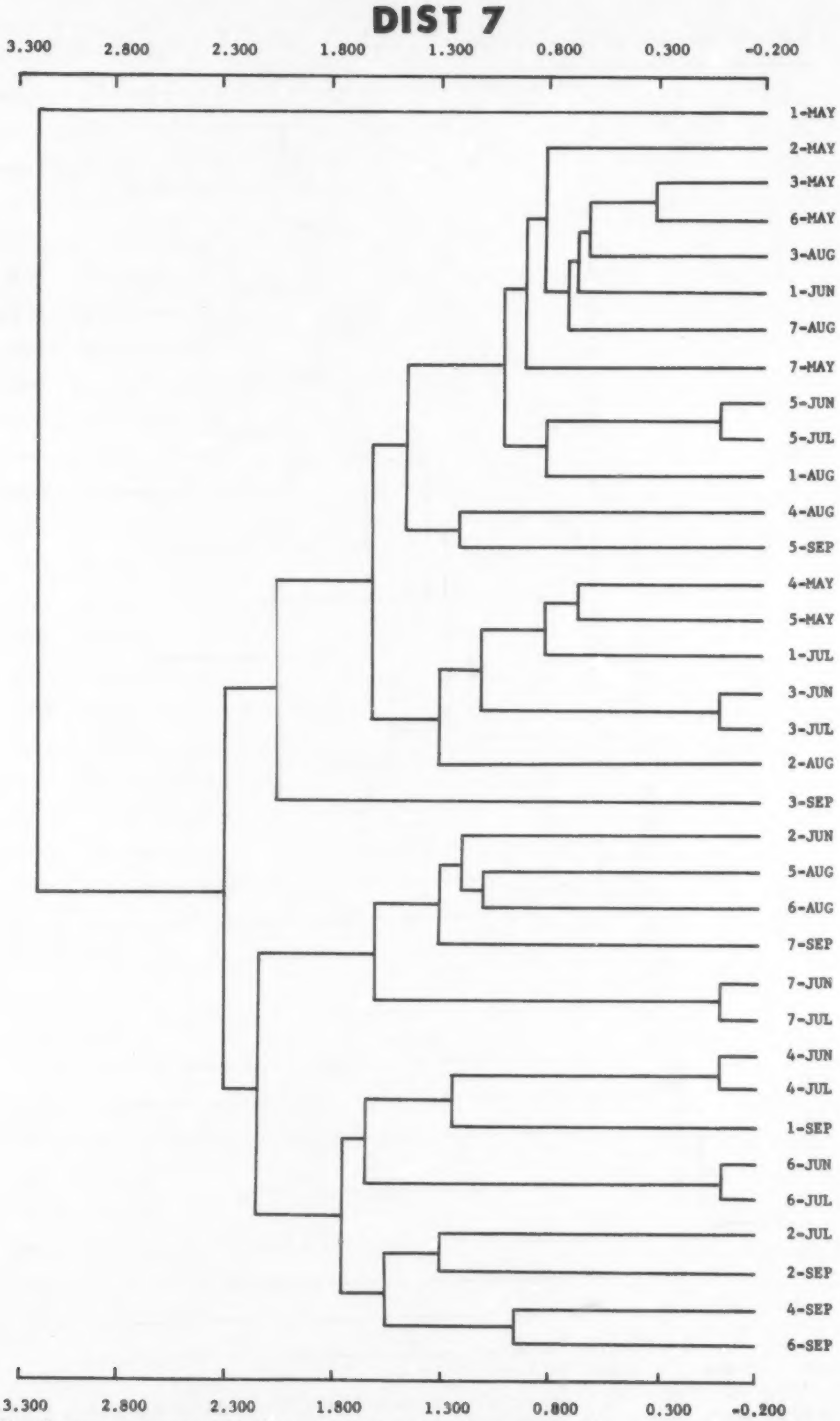


Figure 20. Dendrogram computed from Q-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation and standardized by rows; shows faunal similarities between samples collected from the Cumberland River in 1975.

coefficient (quantitative, distance, square root transformation with standardization by rows) produced the third acceptably low level of distortion.

For the 1973 data set, the principal cause of association (faunal similarity) based on an interpretation of biological and physicochemical data (including cluster analyses interpretation based on the formation of distinct clusters) was the location of the sampling station. In some cases this analysis differed from an analysis based solely on a listing of the highest (i.e., highest similarity level) biological associations in the dendrogram. For the 1975 data set, the principal cause of association based on an interpretation of biological and physicochemical data was in every case different from an interpretation based solely on a listing of the clusters formed at the highest levels of similarity. While these lists identify the most similar pair groups, they do not provide insight into community structure, and for this reason we do not recommend the use of mere similarity-level lists as a basis for interpretaton of biological data.

5.3.2 R-Mode Analysis

5.3.2.1 Introduction--

The purpose of R-mode cluster analysis is to identify recurring groups of species that form biological communities (i.e., associations). If the biological communities are discrete, the cluster analysis will form distinct clusters. If the communities intergrade, the results of cluster analysis will show this.

Whether discrete communities exist has been the subject of numerous ecological studies, and investigators are almost equally divided as to whether their analyses support or contradict the community concept. In a series of studies of rivers of the eastern and southeastern United States, Patrick (1961, 1967) noted that "the number of species of the major groups of organisms--that is, the algae, protozoa, other lower invertebrates, insects, and fish--remain similar from season to season in the same stream. Likewise in similar types of streams they are similar." In other words, the total number of different kinds of organisms is remarkably constant from one system to another. Patrick (1961, 1967) also observed that unstressed, healthy systems usually consist of many species with only a few individuals and a few species with many individuals. Plotting ranked abundance of species against number of species characteristically gives a truncated log-normal curve (Figure 21). Although such generalities do not prove the existence or nonexistence of communities or even associations of species, they indicate similarity among the biotic components of unpolluted streams.

A familiar work suggesting the existence of specific species-environmental associations is the study of indicator organisms developed in Europe by Kolkwitz and Marrson (1908, 1909). In their saprobian system, Kolkwitz and Marrson associated certain species with different zones or regions of water quality. As pointed out by Cairns et al. (1972), this system "was a logical extrapolation of the niche concept (Hutchinson, 1957; Parker and Turner, 1961). That is, each organism has a particular set of environmental prerequisites essential to its survival." An illustration of the type of information resulting from study of indicator organism is shown in Figure 22.

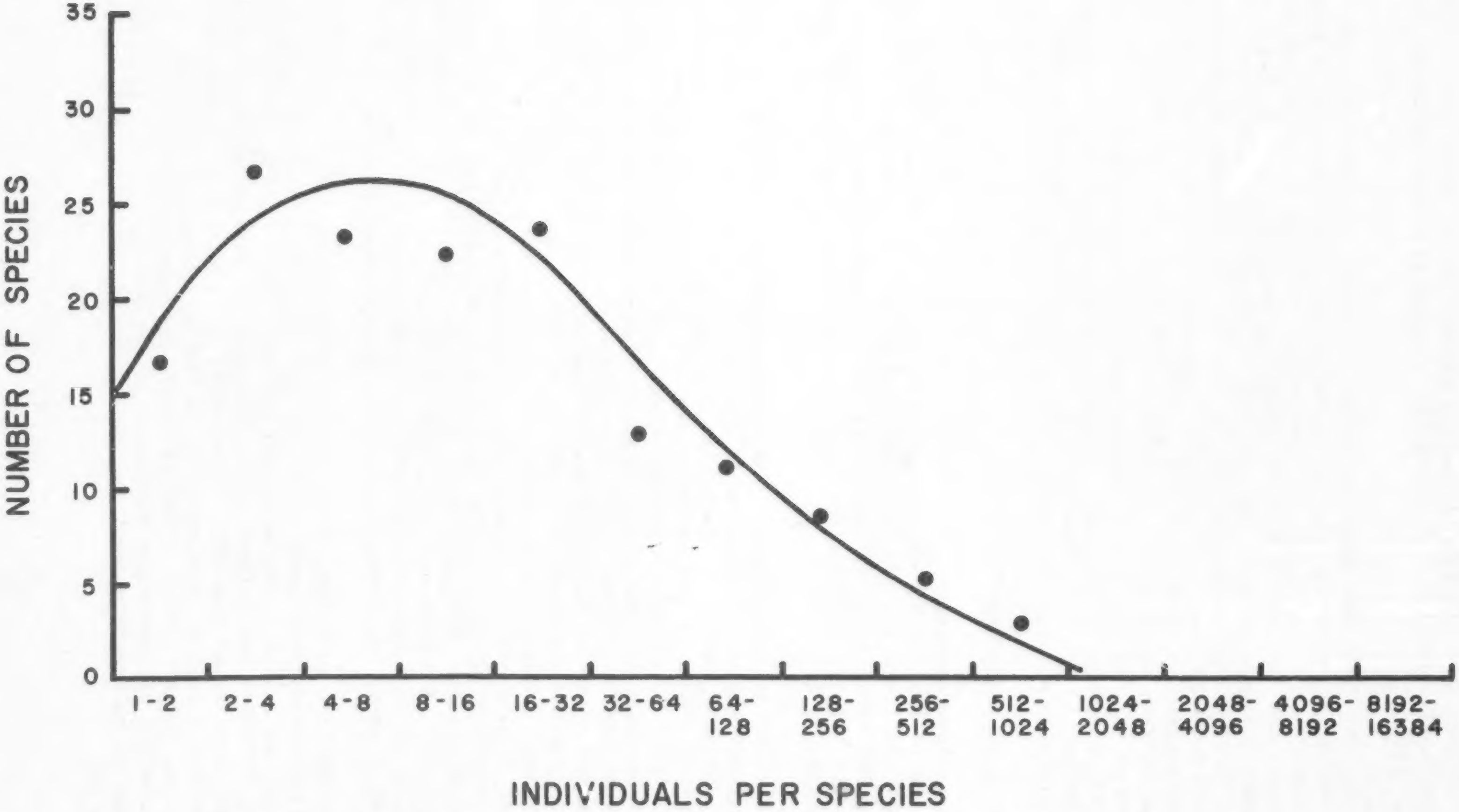


Figure 21. A truncated log-normal distribution fitted to a distribution of species in an aquatic ecosystem not adversely affected by environmental stress.

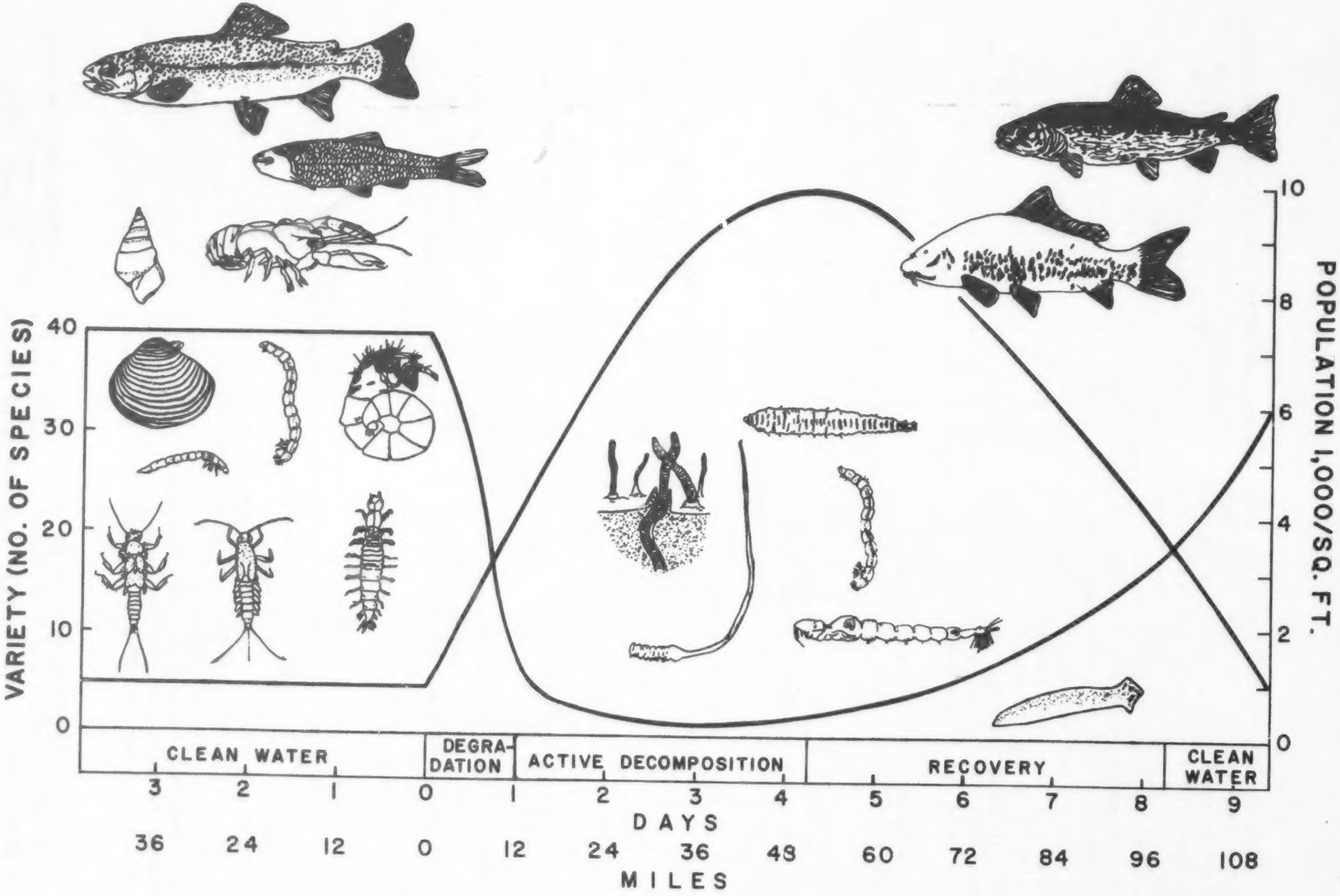


Figure 22. Responses of organisms to severe organic enrichment: changes in types of organisms present, population densities, and biological diversity.

In general, studies indicate that structural pattern exists within freshwater communities. Problems confronting the ecologist are (1) identifying and characterizing species associations and (2) identifying and understanding environmental variables that control the presence of these associations. Stephenson (1972) discussed the first problem in a paper describing the use of computers to classify marine benthic communities. He identified three basic characteristics or attributes that a species must have to belong to an association. The first characteristic was dominance, which is usually expressed as the number of organisms or biomass per unit area. The other characteristics were constancy and fidelity. According to Stephenson (1972), "A species is highly constant if it appears in all the samples or quadrants within an association, but it need not be restricted to a single association. Conversely, a species is highly faithful if it occurs in a single association, but it need not occur within all the samples within the association."

Although the theory of community ecology is based on associations of species, the discussion that follows is based on taxonomy at the generic level or higher. Use of less refined taxonomy is necessitated by the difficulties inherent in large aquatic ecological surveys where taxonomy must be based solely on morphological observations of often limited size classes and sometimes damaged specimens. There also remain, of course, taxonomic uncertainties and discrepancies in certain groups.

5.3.2.2 Clinch River Data Set--

To identify representative assemblages of taxa, R-mode clustering of the Clinch River data was undertaken with six similarity coefficients--S_J, S_{SM}, Corr 6, Corr 7, Dist 6, and Dist 7. Because the complete Clinch River data set included samples collected before and after the pH stress, it was divided into affected and unaffected subsets. R-mode clustering was done on each subset to determine (1) whether assemblages of taxa changed after the spill and (2) whether keeping all 36 zoomacrobenthic samples in the same data matrix changed the clusters of taxa identified in each subset.

The r_{cc} for each dendrogram is listed in Table 27 (refer back to section 5.3.1.1). Because r_{cc} values for Corr 6, Corr 7, and Dist 6 dendrograms were less than 0.8, no discussion of these dendrograms is included here. The r_{cc} values for the S_J, S_{SM}, and Dist 7 dendrograms were greater than 0.8, except for one S_{SM} dendrogram that had a value of 0.76.

To interpret the results, one or two arbitrary levels of similarity were selected for each dendrogram. The clusters or assemblages of taxa found at each level of similarity were tabulated, and the tables compared. For eight dendrograms (Figures 23 to 30) with an r_{cc} of ≥ 0.8, 11 tables were formed, and the dendrograms were grouped according to their overall similarity to each other (Tables 37 through 39). Trophic, functional (TF) codes (refer to Table 12) were then assigned to each taxon, and intercluster reordering by numeric code was initiated. The TF clusters were then rearranged within each dendrogram to show similarities among dendrograms (Tables 40 to 42). The last step was to decode the TF clusters and list by scientific name (Tables 43 to 45).

5.3.2.2.1 Presence-absence data--In Table 39, four clusters of taxa were identified at the 0.87 and 0.89 levels of similarity for the S_J and S_{SM} dendrograms computed from samples unaffected by the spill. Cluster 1 had eight taxa, dominated by pH-tolerant beetles. Two major trophic-functional

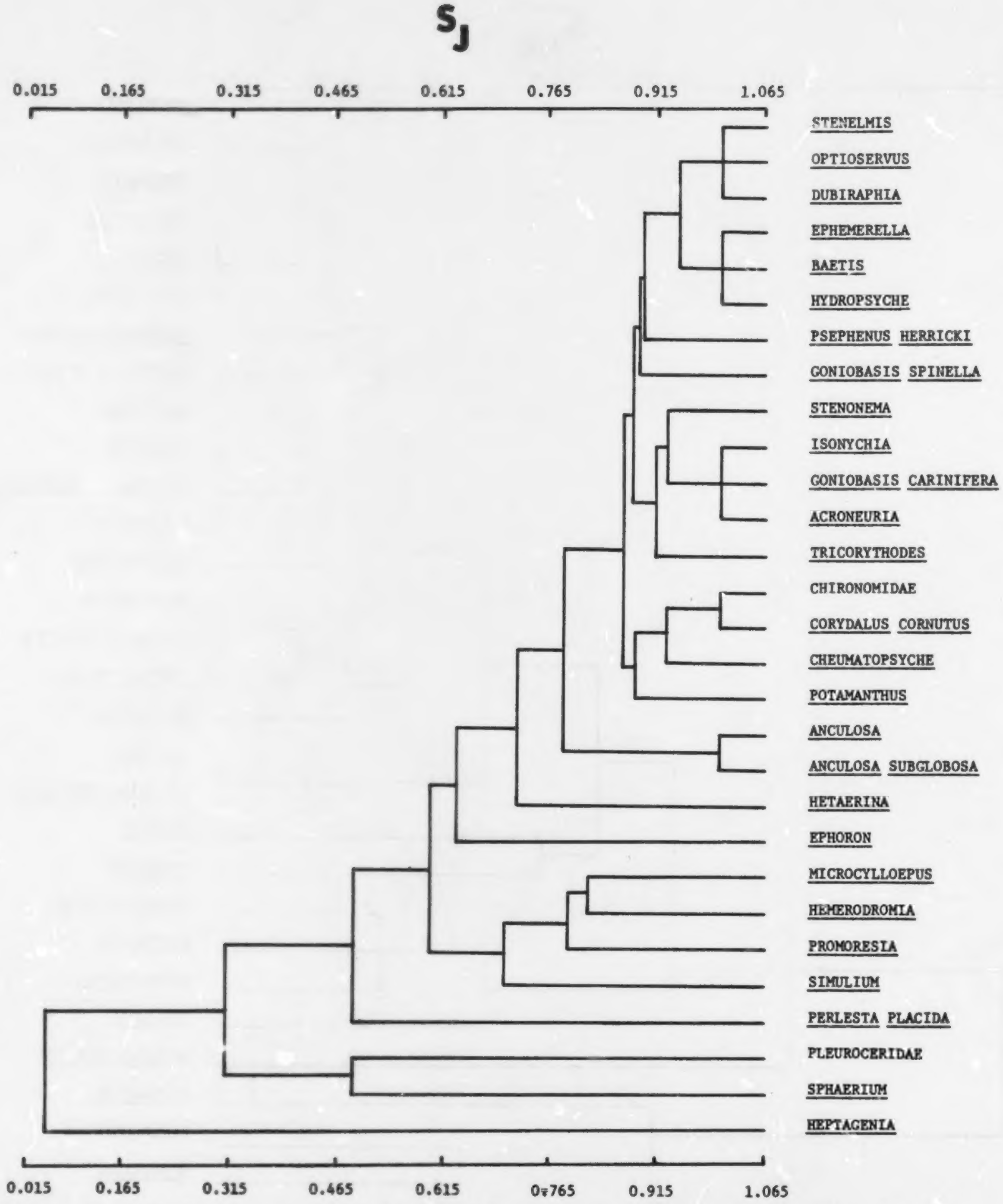


Figure 23. Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from stations on the Clinch River unaffected by low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.97).

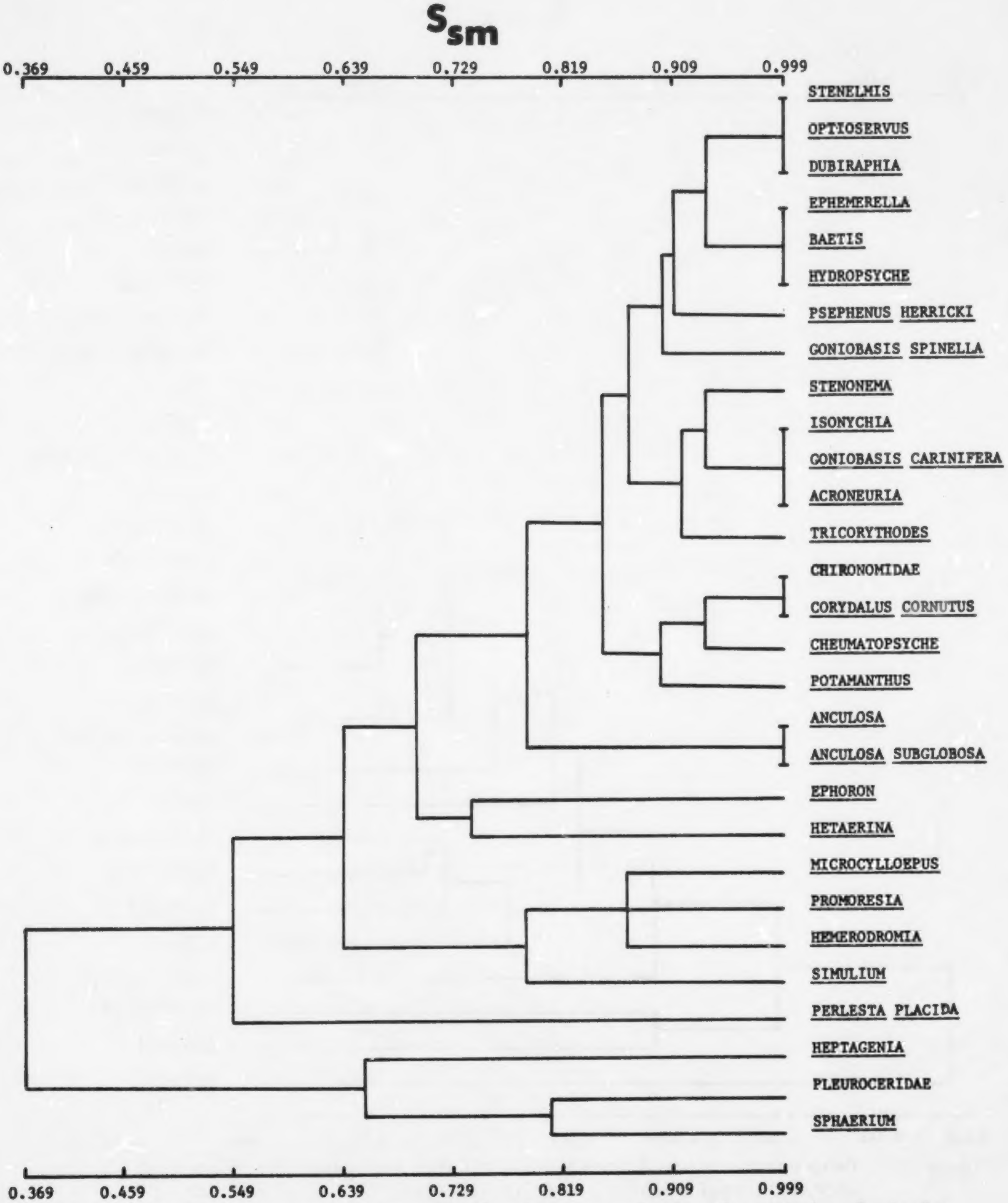


Figure 24. Dendrogram computed from R-mode cluster analysis of a matrix of simple matching coefficients, showing distributional similarities of taxa collected from stations on the Clinch River unaffected by low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.91).

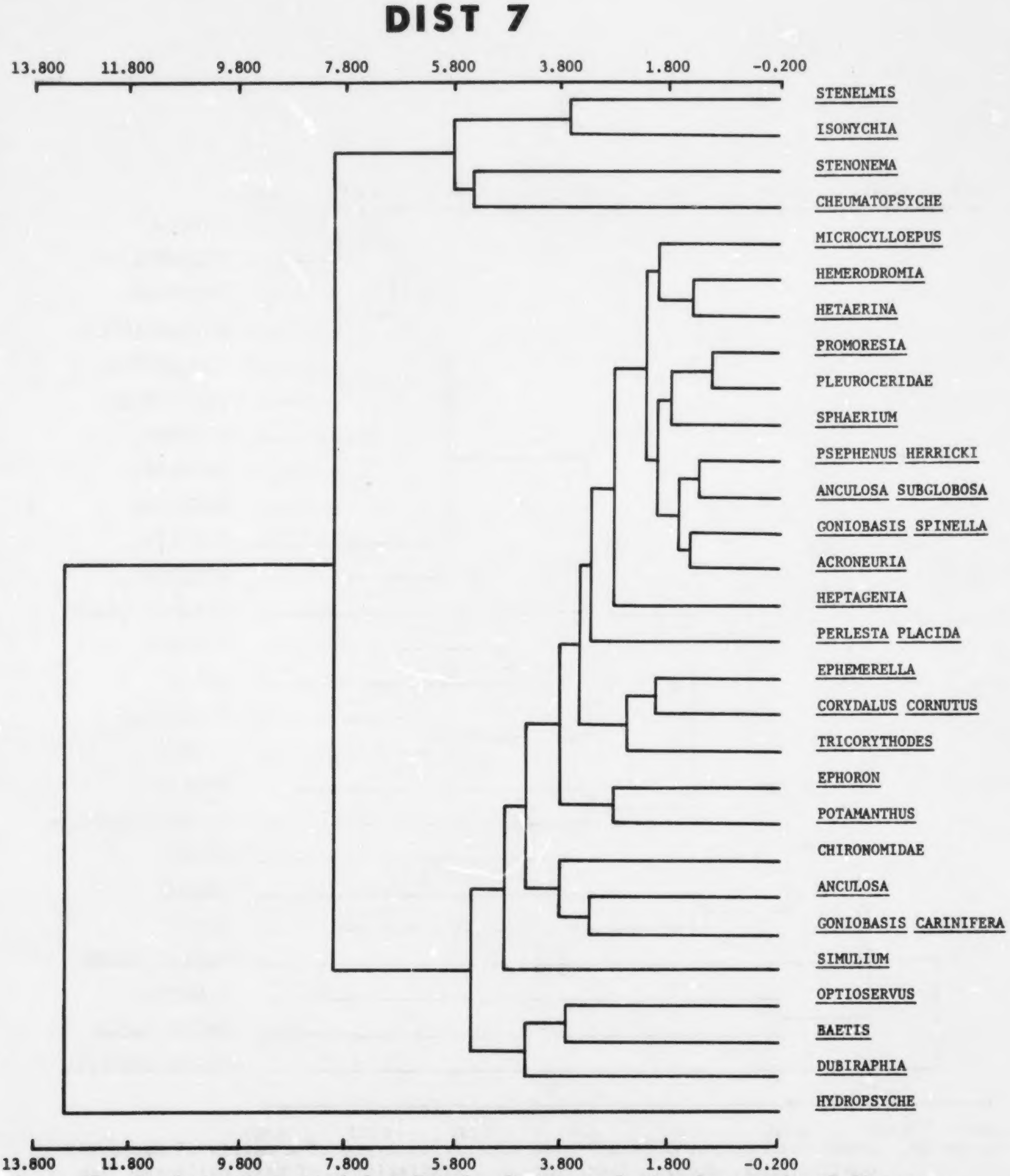


Figure 25. Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation, showing distributional similarities of taxa collected from stations on the Clinch River unaffected by low-pH stress that resulted from the 1970 spill of acid ($r_{cc} = 0.92$).

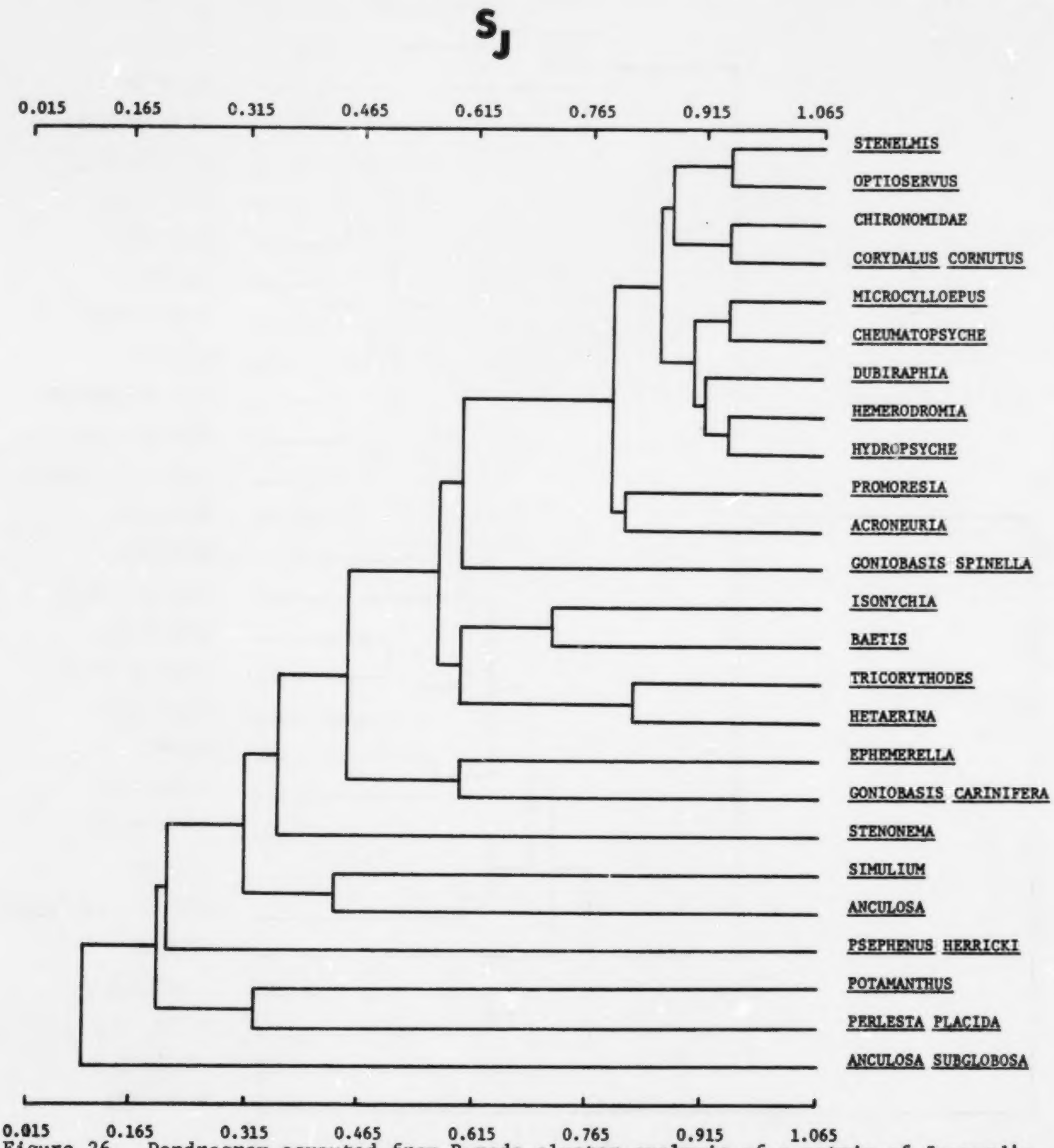


Figure 26. Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from stations on the Clinch River affected by low-pH stress that resulted from from the 1970 spill of acid (r_{cc} = 0.97).

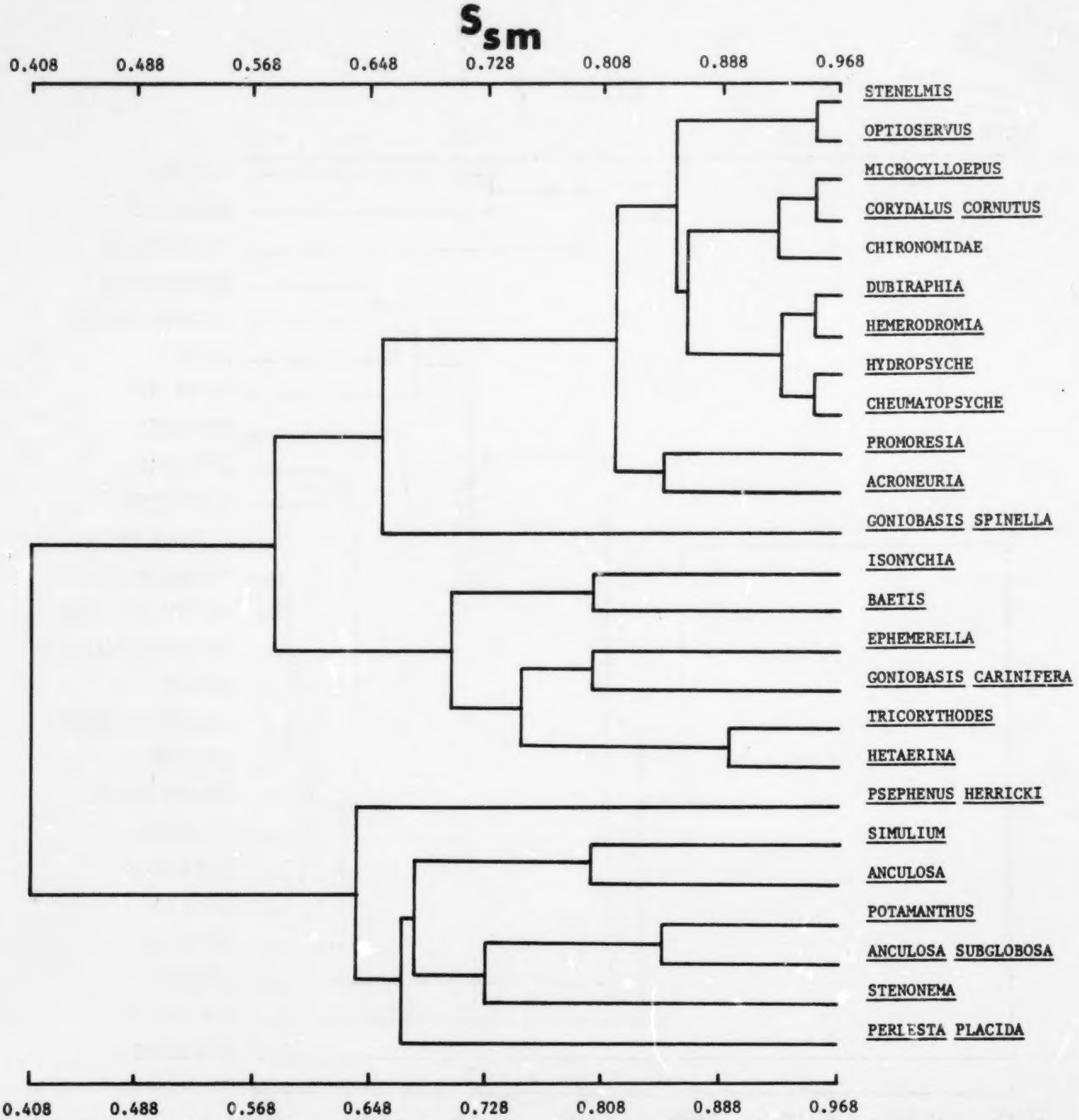


Figure 27. Dendrogram computed from R-mode cluster analysis of a matrix of simple matching coefficients, showing distributional similarities of taxa collected from stations on the Clinch River affected by low-pH stress that resulted from the 1970 spill of acid ($r_{cc} = 0.84$).

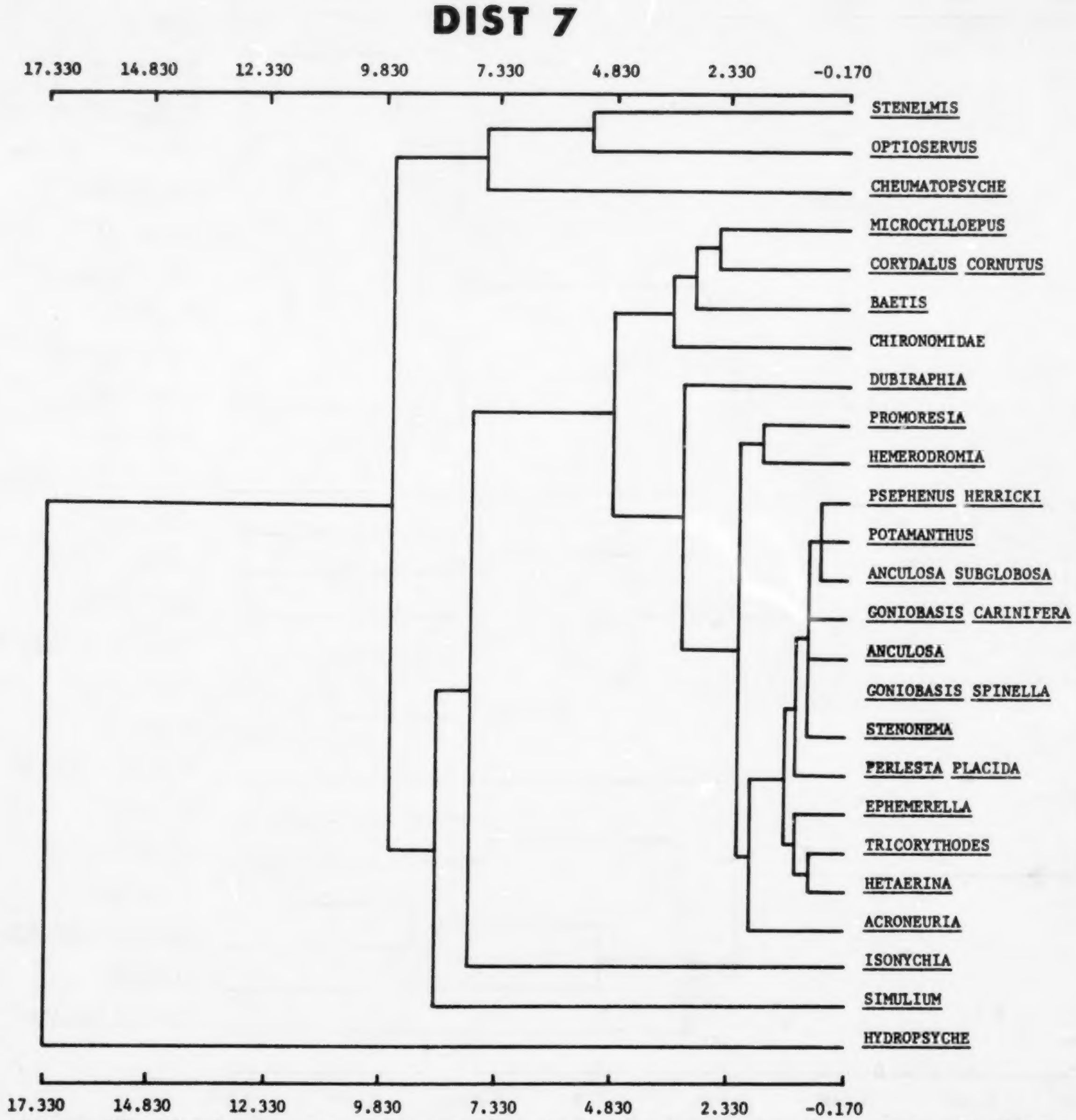
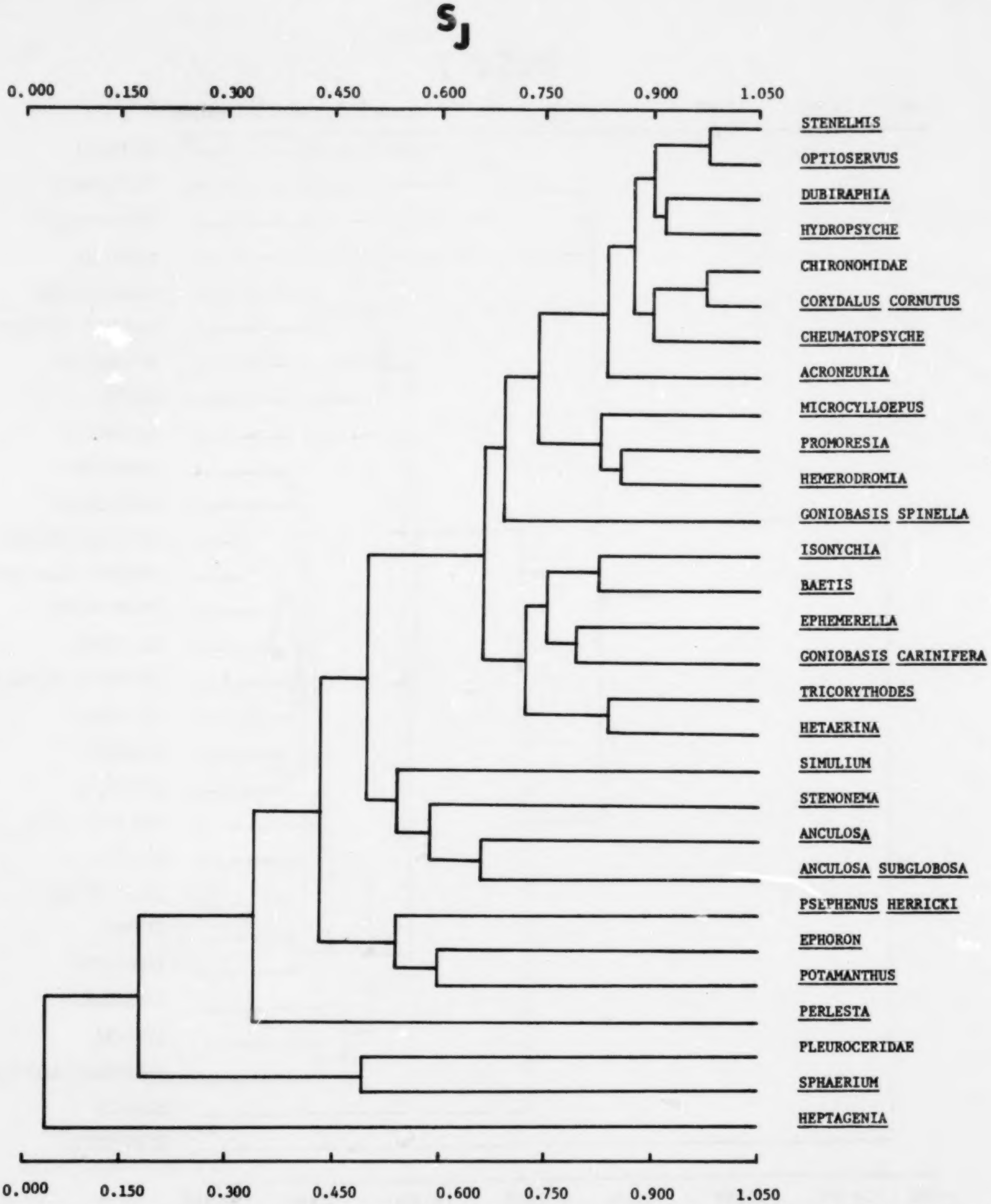


Figure 28. Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation, showing distributional similarities of taxa collected from stations on the Clinch River affected by low-pH stress that resulted from the 1970 spill of acid ($r_{cc} = 0.97$).



S_J
0.000
0.150
0.300
0.450
0.600
0.750
0.900
1.050
STENELMIS
OPTIOSERVUS
DUBIRAPHIA
HYDROPSYCHE
CHIRONOMIDAE
CORYDALUS CORNUTUS
CHEUMATOPSYCHE
ACRONEURIA
MICROCYLLOEPUS
PROMORESIA
HEMERODROMIA
GONIOBASIS SPINELLA
ISONYCHIA
BAETIS
EPHEMERELLA
GONIOBASIS CARINIFERA
TRICORYTHODES
HETAERINA
SIMULIUM
STENONEMA
ANCULOSA
ANCULOSA SUBGLOBOSA
PSEPHENUS HERRICKI
EPHORON
POTAMANTHUS
PERLESTA
PLEUROCERIDAE
SPHAERIUM
HEPTAGENIA
0.000
0.150
0.300
0.450
0.600
0.750
0.900
1.050

Figure 29. Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from stations on the Clinch Riber both affected and unaffected by the low-pH stress that resulted from the 1970 spill of acid (r_{cc} = 0.95). Only those taxa are included that comprise 10 percent or more of the total samples.

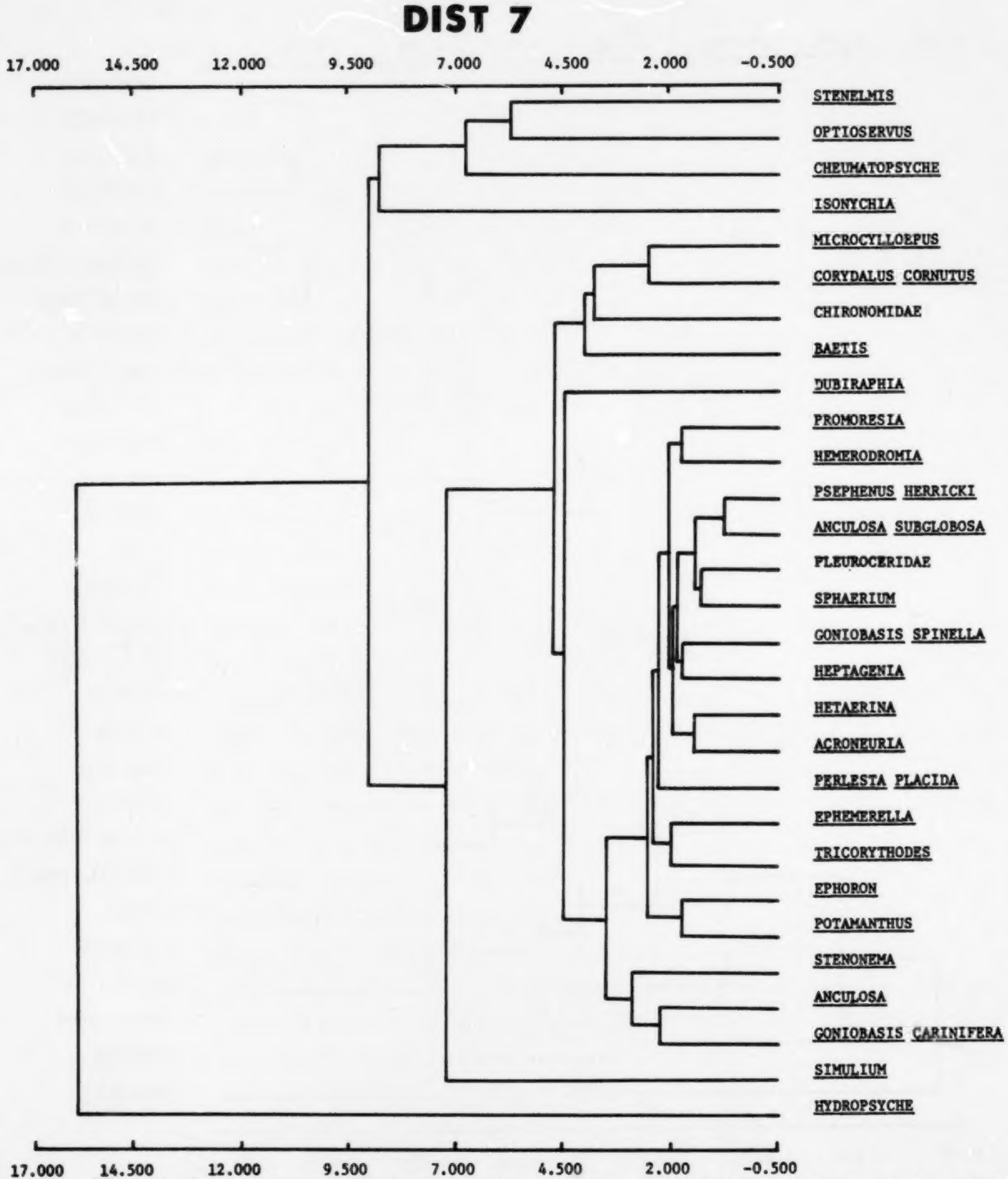


Figure 30. Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients computed from data that had been transformed by the square-root transformation, showing distributional similarities of taxa collected from stations on the Clinch River both affected and unaffected by the low-pH stress that resulted from the 1970 spill of acid ($r_{cc} = 0.97$). Only those taxa are included that comprise 10 percent or more of the total sample.

groups were represented--collectors and grazers. Clusters 2 and 3 had four and five taxa, respectively, but each cluster had three major functional groups--collectors, grazers, and predators. This indicated that assemblages of taxa represented in clusters 2 and 3 were more characteristic of the total zoomacrobenthic community than those in cluster 1. The fourth cluster in Table 39 consisted of two snail taxa representing one TF group.

Table 43 shows that several taxa clustered together (using presence-absence coefficients), regardless of the level of similarity, type of data considered, or size of the data matrix: (1) *Optioservus*-*Stenelmis*, (2) Chironomidae-*Corydalus ornutus*, (3) *Hydropsyche*-*Cheumatopsyche*-*Dubiraphia*, (4) *Isonychia*-*Baetis*, (5) *Goniobasis carinifera*-*Ephemerella*, and (6) *Tricorythodes*-*Hetaerina*. These were identified as (1) grazers, (2) grazer-predators, (3) collector-grazers, (4) collector-grazers, (5) grazers, and (6) collector-predators.

The six clusters were evaluated with regard to constancy and fidelity.

5.3.2.2.2 Quantitative data, species counts--Associations of taxa based on counts of individuals per taxon were more difficult to identify and less discrete. Table 44 summarizes clusters of taxa computed with the distance coefficient Dist 7. Only three taxa (*Psephenus herricki*, *Goniobasis spinella*, and *Hetaerina*) met the criteria of constancy and fidelity. Unfortunately, the information provided by this coefficient was of limited value, and other quantitative techniques should be considered in future studies.

5.3.2.2.3 Summary--One justification for dividing the original data matrix into subsets was to determine whether the associations of taxa found at stations unaffected by the spill were different from those found at stations affected by the spill. To answer this question, the associations identified in Tables 39 and 43 were collated. Table 45 shows that species associations after the spill were much smaller (two to three taxa per association), than before the spill. The table also shows that *Stenelmis*-*Optioservus* and Chironomidae-*Corydalus cornutus* were the only associations found in dendrograms of both unaffected and affected stations. This indicates that associations were influenced by the pH stress and that division of the original data matrix of 36 samples into subsets was justified.

5.3.2.3 Cumberland River Data Set--1973--

5.3.2.3.1 Presence-absence data--Jaccard's coefficient S_J produced clusters with low levels of similarity (Figure 31). At a similarity level of 0.27, four clusters were formed, with several taxa left unclustered. The first cluster consisted of a heterogeneous association of nine taxa from three phyla--five detrital collectors, three predators, and one grazer. This group included one omnivore and almost equal numbers of detritivores and carnivores. Cluster 2 contained a small but equally heterogeneous group of taxa displaying three different food habits. The last two clusters consisted entirely of wormlike taxa that are usually abundant only in streams that have been organically enriched.

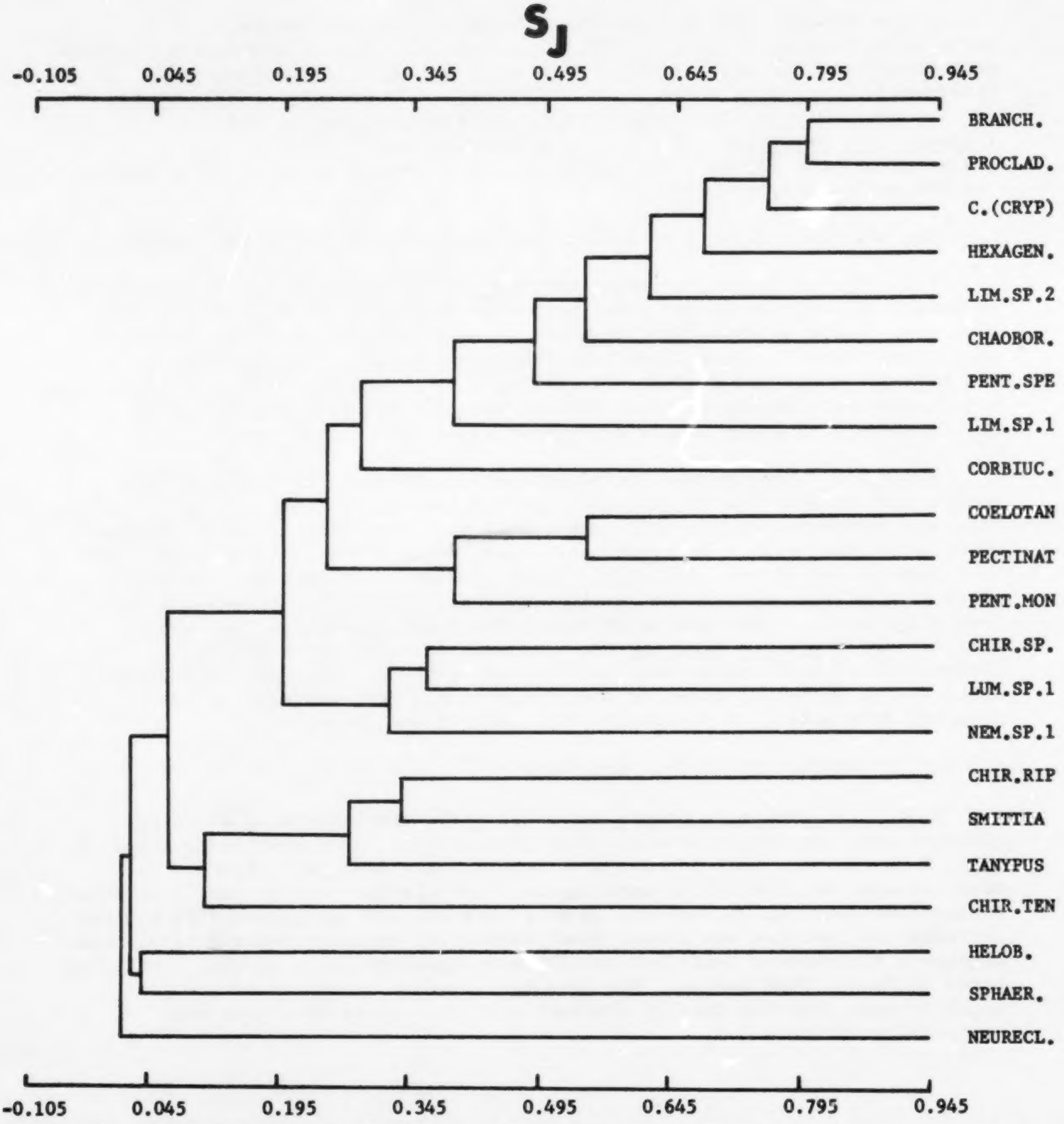


Figure 31. Dendrogram computed from R-mode cluster analysis of a matrix of Jaccard's coefficients, showing distributional similarities of taxa collected from the Cumberland River in 1973.

Cluster analysis of the simple matching coefficient S_{SM} produced clusters with low levels of similarities (Figure 32). Two clusters were formed at a level of similarity of 0.75. The first cluster comprises four taxa that were also closely clustered with Jaccard's coefficient S_J. The second cluster is a heterogeneous association of taxa that had been unclustered or placed in small clusters by Jaccard's coefficient.

5.3.2.3.2 Quantitative data, species counts--The distance coefficient Dist 6 (square-root transformation of data and standardization by rows) produced clusters with low levels of similarity (Figure 33). Seven small clusters were formed at a level of similarity of 1.11, with six taxa remaining unclustered. Only one association had also occurred when presence-absence coefficients were used.

The distance coefficient Dist 7 (square-root transformation of data) produced a dendrogram in which almost all the taxa were grouped at the same level of similarity and thus contained little useful information about discrete associations (Figure 34).

The correlation coefficients Corr 6 and Corr 7 produced identical matrices and dendrograms due to the standardization inherent in the use of the product-moment correlation coefficient. Again, levels of similarity of clusters were fairly low. At levels of 0.46 and 0.35, five and seven small clusters, respectively, were formed (Figure 35). Some clusters contained functionally and morphologically similar taxa, but others did not. Some contained only frequently collected taxa, whereas others contained mixtures of frequent and rare taxa. Very little ecological information that pertained to the question of associations of taxa came from study of these dendrograms.

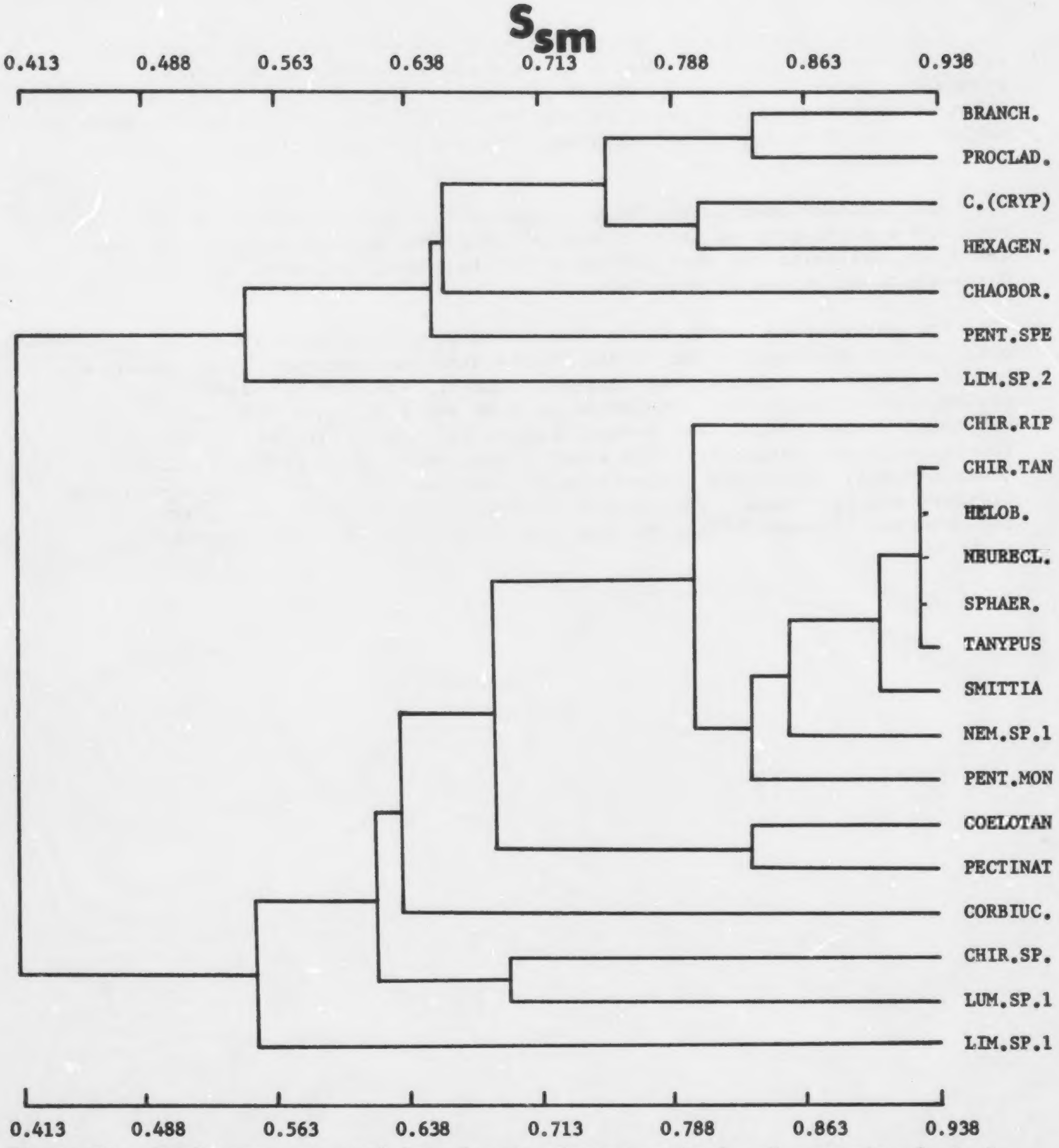


Figure 32. Dendrogram computed from R-mode cluster analysis of a matrix of simple matching coefficients, showing distributional similarities of taxa collected from the Cumberland River in 1973.

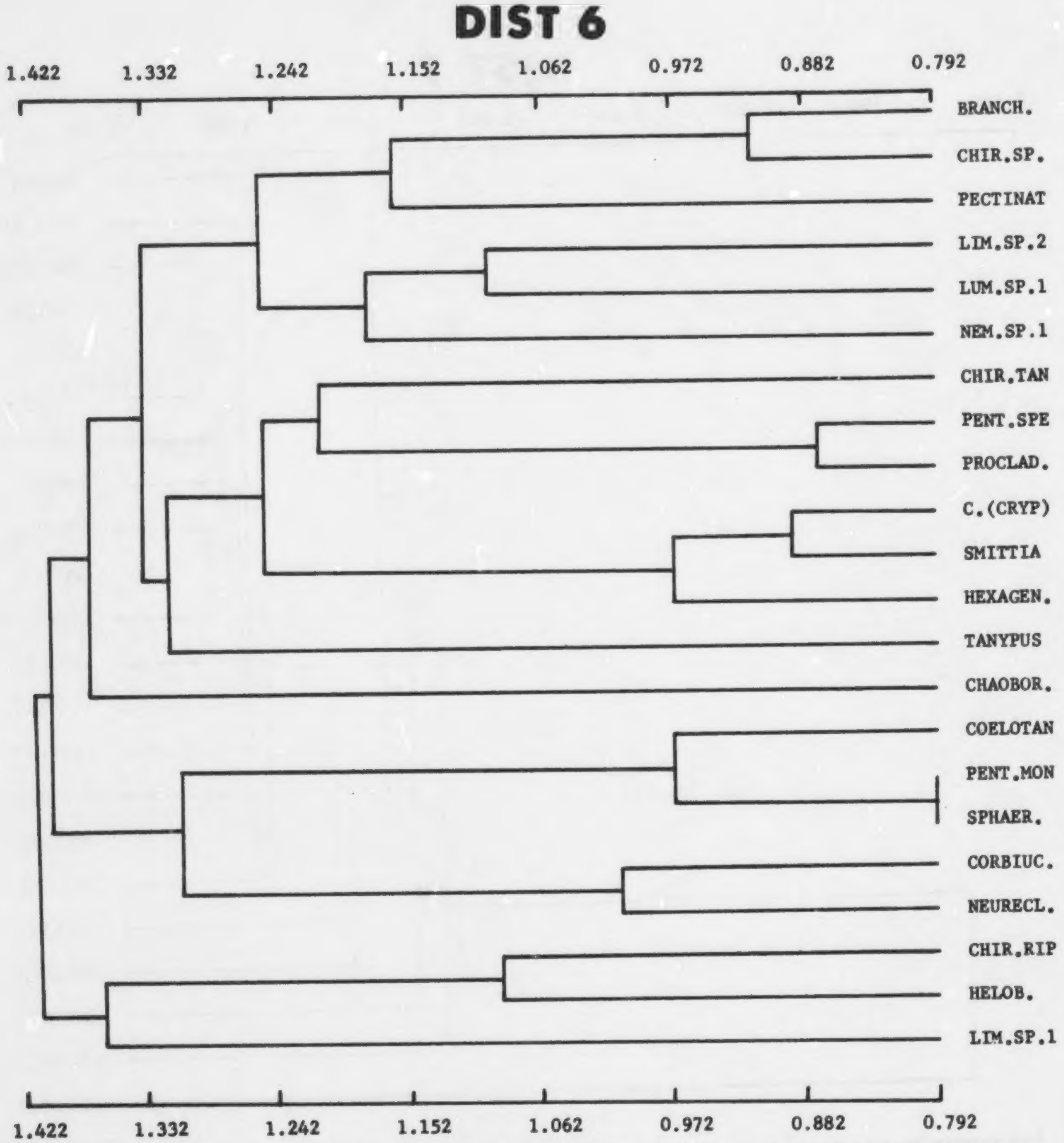


Figure 33. Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients with data transformed with the square-root transformation and standardized by rows, showing distributional similarities of taxa collected from the Cumberland River in 1973.

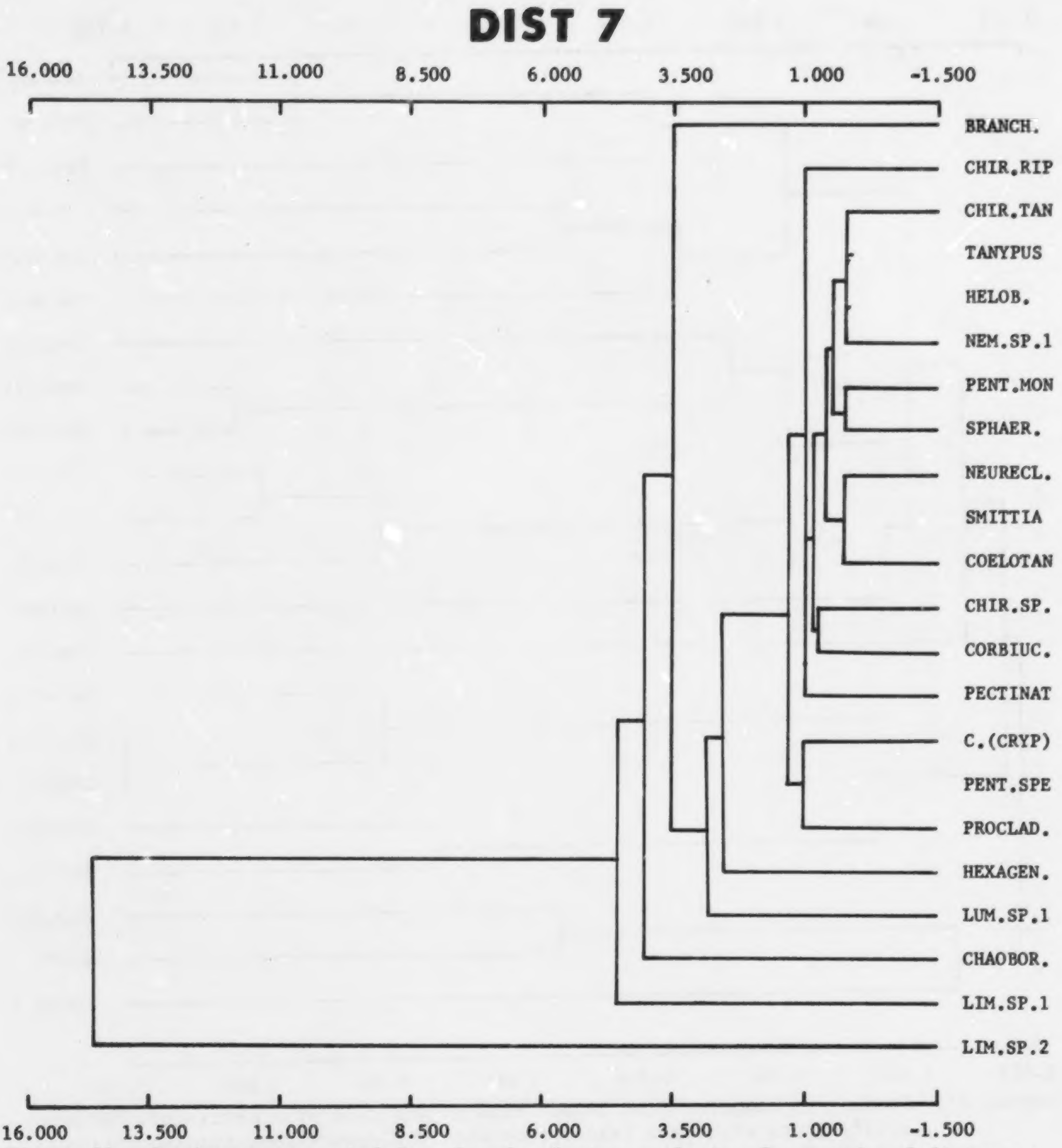
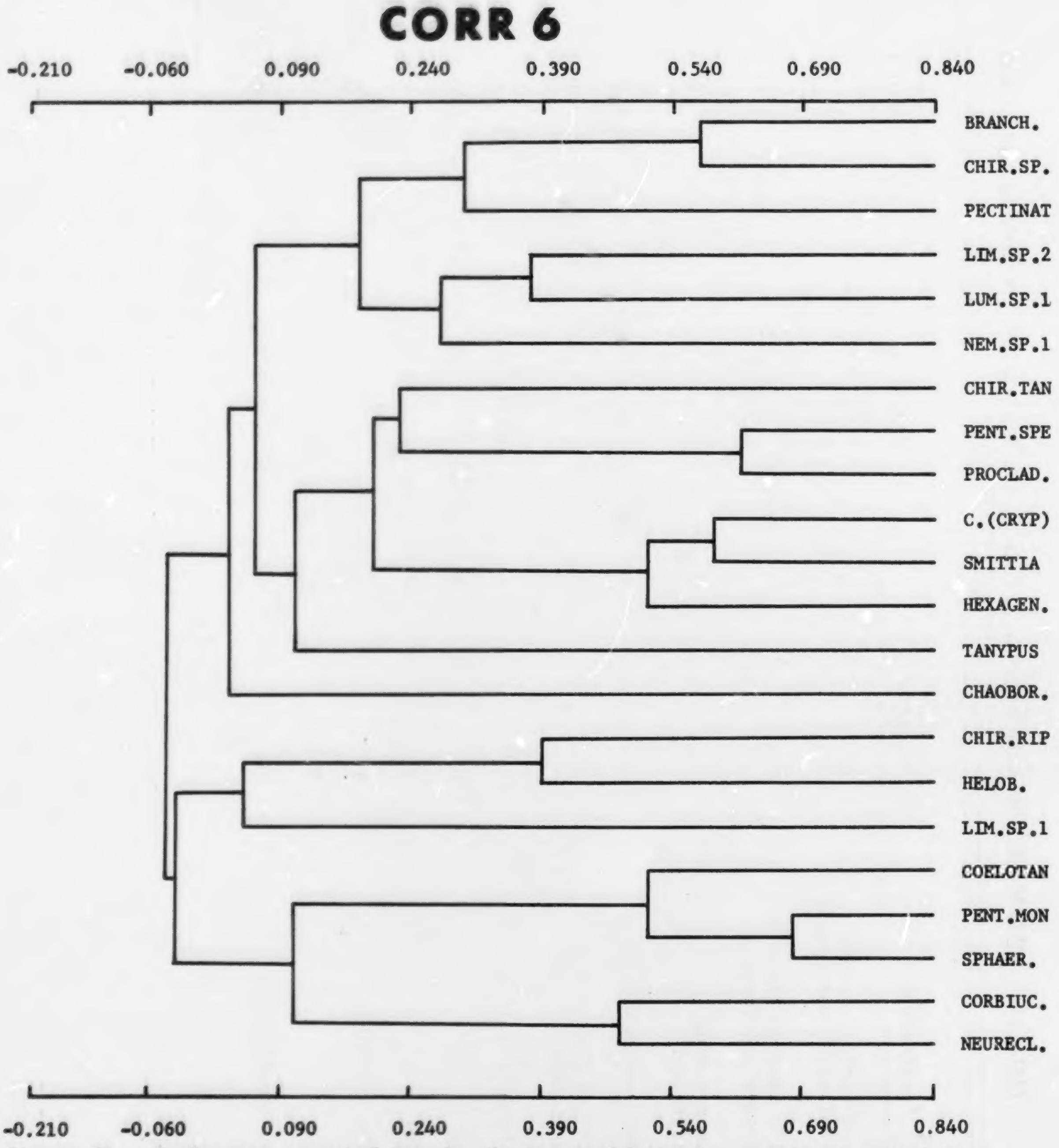


Figure 34. Dendrogram computed from R-mode cluster analysis of a matrix of distance coefficients with data transformed by the square-root transformation, showing distributional similarities of taxa collected from the Cumberland River in 1973.



CORR 6

Figure 35. Dendrogram computed R-mode cluster analysis of a matrix of correlation coefficients with data transformed by the square-root transformation and standardized by rows, showing distributional similarities of taxa collected from the Cumberland River in 1973.

TABLE 16. COEFFICIENTS OF CORRELATION, DISTANCE, AND SIMILARITY: ABBREVIATIONS, EQUATIONS, AND UPPER AND LOWER LIMITS[a]

Coefficient	Abbreviation	Equation	Lower limit	Upper limit
Pearson product-moment correlation coefficient	Corr	$r_{jk} = s_{jk}/s_j s_k$	-1	1
Sokal's average taxonomic distance coefficient	Dist	$d_{jk} = \sqrt{\{\Sigma(X_{ij} - X_{ik})2/n\}}$	0	∞
Simple matching coefficient	S_{SM}	$S_{SM} = (a + d)/n$	0	1
Unnamed coefficient 1	UN1	$S_{UN1} = 2(a + d)/n + a + d$	0	1
Rogers and Tanimoto's coefficient	RT	$S_{RT} = (a + d)/(a + d + 2b + 2c)$	0	1
Unnamed coefficient 3	UN3	$S_{UN3} = (a + d)/(b + c)$	0	∞
Hamann	H	$S_H = (a + d - b - c)/n$	0	1
Jaccard	S_J	$S_J = a/(a + b + c)$	0	1
Russell and Rao	RR	$S_{RR} = a/n$	0	1
Dice	D	$S_D = 2a/(2a + b + c)$	0	1
Unnamed coefficient 2	UN2	$S_{UN2} = a/(a + 2b + 2c)$	0	1

TABLE 16 (continued)

Coefficient	Abbreviation	Equation	Lower limit	Upper limit
Kulczynski 1	K1	$S_{K1} = a/(n_j + n_K - 2a)$	0	∞
Kulczynski 2	K2	$S_{K2} = 1/2(a/n_j) + 1/2(a/n_K)$	0	1
Unnamed coefficient 4	UN4	$S_{UN4} = 1/4(a/n_J) + 1/4(a/n_K)$	0	1
Ochiai (Otsuka)	OCH	$S_{OCH} = a/\sqrt{(n_J\ n_K)}$	0	1
Unnamed coefficient 5	UN5	$S_{UN5} = ad/\sqrt{(n_J + n_K + n_j + n_k)}$	0	1
Yule	Y	$S_Y = (ad - bc)/(ad + bc)$	-1	1
Phi	PHI	$S_{PHI} = (ad - bc)/\sqrt{(n_J\ n_K\ n_j\ n_k)}$	-1	1

[a] See Sokal and Sneath (1963) for further discussion.

TABLE 17. CONTINGENCY TABLE (2 X 2) DEFINING THE TERMS *a*, *b*, *c*, AND *d* AS USED IN THE EQUATIONS IN TABLE 13

	Sample j	
Sample k	Species present	Species absent
Species present	a (present in both)	b (present in k, absent from j)
Species absent	c (present in j, absent from k)	d (absent from both; negative match)

TABLE 18. EFFECT OF THE TRANSFORMATION LOG (X_{ij} + 1), WHERE X IS THE ABUNDANCE OF THE *i*th SPECIES IN THE *j*th SAMPLE

Abundance	Value after transformation
0	0
9	1
99	2
999	3
9999	4

TABLE 19. EFFECT OF THE TRANSFORMATION $\sqrt{(X_{ij} + 0.5)}$, WHERE X_{ij} IS THE ABUNDANCE OF THE *i*th SPECIES IN THE *j*th SAMPLE

Abundance	Value after transformation
0	0.707
10	3.240
100	10.025
1000	31.631
10000	100.003

TABLE 20. LABELS OF CORRELATION AND DISTANCE MATRICES WITH VARIOUS TRANSFORMATION

Transformation	Correlation	Distance
None (raw data)	Corr 1	Dist 1
Standardization by rows	Corr 2	Dist 2
Log transformation, standardization by rows	Corr 4	Dist 4
Square root transformation, standardization by rows	Corr 6	Dist 6
Square root transformation, no standardization	Corr 7	Dist 7

TABLE 21. MATRIX OF COEFFICIENTS OF COPHENETIC CORRELATION COMPUTED BETWEEN CORRESPONDING ELEMENTS OF 21 CORRELATION AND SIMILARITY MATRICES[a]

	Corr 1	Corr 2	Corr 4	Corr 6	Corr 7	S_{SM}	UN1	RT	UN3	H	S_J
Corr 1	1.000										
Corr 2	0.021	1.000									
Corr 4	0.137	0.786	1.000								
Corr 6	0.086	0.925	0.943	1.000							
Corr 7	0.988	0.055	0.202	0.139	1.000						
S_{SM}	0.121	0.494	0.646	0.571	0.171	1.000					
UN1	0.116	0.478	0.622	0.550	0.165	0.996	1.000				
RT	0.127	0.509	0.668	0.590	0.176	0.994	0.981	1.000			
UN3	0.129	0.446	0.600	0.525	0.163	0.759	0.718	0.809	1.000		
H	0.121	0.494	0.646	0.571	0.171	1.000	0.996	0.994	0.759	1.000	
S_J	0.224	0.283	0.537	0.421	0.282	0.700	0.684	0.712	0.599	0.700	1.000
RR	0.223	0.138	0.348	0.257	0.273	0.424	0.418	0.427	0.338	0.424	0.916
D	0.240	0.253	0.508	0.389	0.300	0.668	0.658	0.673	0.532	0.668	0.987
UN2	0.211	0.302	0.554	0.441	0.266	0.711	0.689	0.732	0.662	0.711	0.989
K1	0.169	0.286	0.507	0.409	0.208	0.611	0.577	0.654	0.805	0.611	0.804
K2	0.336	0.237	0.531	0.386	0.407	0.648	0.636	0.654	0.521	0.648	0.900
UN4	0.289	0.348	0.631	0.491	0.360	0.828	0.817	0.832	0.651	0.828	0.858
OCH	0.282	0.253	0.529	0.397	0.348	0.674	0.663	0.679	0.538	0.674	0.972
UN5	0.251	0.339	0.618	0.484	0.316	0.812	0.796	0.822	0.674	0.812	0.958
Y	0.344	0.211	0.486	0.341	0.415	0.619	0.619	0.608	0.417	0.619	0.678
PHI	0.270	0.351	0.637	0.498	0.341	0.834	0.832	0.837	0.657	0.834	0.895

[a] Abbreviations defined in Table 20.

TABLE 21 (continued)

	RR	D	UN2	K1	K2	UN4	OCH	UN5	Y	PHI
Corr 1										
Corr 2										
Corr 4										
Corr 6										
Corr 7										
S_{SM}										
UN1										
RT										
UN3										
H										
S_J										
RR	1.000									
D	0.910	1.000								
UN2	0.889	0.955	1.000							
K1	0.705	0.731	0.871	1.000						
K2	0.813	0.916	0.870	0.669	1.000					
UN4	0.667	0.856	0.845	0.697	0.951	1.000				
OCH	0.889	0.987	0.940	0.719	0.968	0.911	1.000			
UN5	0.787	0.953	0.943	0.765	0.941	0.957	0.968	1.000		
Y	0.546	0.721	0.633	0.445	0.906	0.912	0.809	0.811	1.000	
PHI	0.718	0.891	0.880	0.707	0.951	0.995	0.933	0.975	0.886	1.000

TABLE 22. COEFFICIENTS OF COPHENETIC CORRELATION COMPARING DISTANCE MATRICES AND SELECTED CORRELATION AND SIMILARITY MATRICES[a]

	Dist 1	Dist 2	Dist 4	Dist 6	Dist 7
Dist 1	1.000				
Dist 2	0.332	1.000			
Dist 4	0.227	0.878	1.000		
Dist 6	0.326	0.978	0.951	1.000	
Dist 7	0.890	0.416	0.320	0.431	1.000
Corr 6	---	---	---	---	---
Corr 7	---	---	---	-0.276	-0.514
S_{SM}	-0.157	-0.412	-0.680	-0.532	-0.190
S_J	0.093	-0.100	-0.358	-0.210	-0.060

[a] Data collected from Cumberland River (Old Hickory Reservoir) in 1973.

TABLE 23. NUMBER OF TAXA IN EACH MAJOR TAXONOMIC GROUP IN THE CLINCH RIVER (1970) BEFORE AND AFTER RELATIVE SPECIES ABUNDANCE WAS DETERMINED AND RARE TAXA WERE ELIMINATED

Major taxonomic group	Number of taxa per group	
	All taxa	Rare taxa eliminated[a]
Amphipoda	1	
Annelida	3	
Coleoptera	22	6
Decapoda	4	
Diptera	13	3
Ephemeroptera	15	8
Gastropoda	14	5
Hemiptera	1	
Hydracarina	1	
Lepidoptera	2	
Megaloptera	3	1
Odonata	14	1
Pelecypoda	6	1
Plecoptera	7	2
Trichoptera	17	2
Total	123	29

[a] Based on a relative abundance (RA) ≤ 0.01.

TABLE 24. TWENTY-NINE TAXA AND THEIR RESPECTIVE TROPHIC CODES FOR THE REDUCED CLINCH RIVER DATA SET, 1970

Taxon	Trophic code*	Taxon	Trophic code*
Stenelmis sp.	3411	*Tricorythodes* sp.	2312
Microcylloepus sp.	3411	*Baetis* sp.	3424
Optioservus sp.	3411	*Anculosa* sp.	3411
Dubiraphia sp.	3411	*Anculosa subglobosa*	3411
Promoresia sp.	3411	Pleuroceridae	3411
Psephenus herricki	3412	*Goniobasis spinella*	3411
Chironomidae	3524	*Goniobasis carinifera*	3411
Simulium sp.	2212	*Corydalus cornutus*	4613
Hemerodromia sp.	4713	*Hetaerina* sp.	4613
Ephoron sp.	2612	*Sphaerium* sp.	2212
Potamanthus sp.	2612	*Perlesta placida*	4624
Stenonema sp.	3424	*Acroneuria* sp.	4624
Heptagenia sp.	3424	*Hydropsyche* sp.	2224
Isonychia sp.	2224	*Cheumatopsyche* sp.	2224
Ephemerella sp.	3424		

*Trophic codes are listed in Tables 11 and 12 (section 4.1.3).

TABLE 25. NUMBER OF TAXA WITH A RELATIVE ABUNDANCE ≥0.01 DIVIDED BY THE TOTAL NUMBER OF TAXA PER STATION, CLINCH RIVER, 1970

Station number	Early June	Late June	Early July	Late July	Early August	Late August
4	28/64			27/72		25/57
7	24/54	24/46	23/38	24/43	20/43	25/50
8	24/47	16/28	16/25	16/30	19/33	15/29
9	25/48	12/24	13/30	17/36	16/29	18/35
10	26/50	19/39	19/35	20/39	22/48	21/43
11	26/49			22/56		23/44

TABLE 26. TAXA WITH A RELATIVE ABUNDANCE ≥0.01 AS PERCENT OF THE TOTAL NUMBER OF TAXA PER STATION, CLINCH RIVER, 1970

Station number	Early June	Late June	Early July	Late July	Early August	Late August
4	44			38		44
7	44	52	61	56	47	50
8	51	57	64	53	58	52
9	52	50	43	47	55	51
10	52	49	54	51	46	49
11	53			39		

TABLE 27. TOTAL NUMBER OF ORGANISMS PER STATION, CLINCH RIVER, 1970

Station number	Early June	Late June	Early July	Late July	Early August	Late August
4	2693			3847		2522
7	1830	853	989	1636	562	912
8	1465	703	901	2618	656	397
9	1738	936	1191	5042	1010	812
10	3857	1186	2189	5614	2945	2351
11	1645			2723		1557

TABLE 28. TWENTY-NINE TAXA WITH RELATIVE ABUNDANCE ≥ 0.01 AS PERCENT OF TOTAL NUMBER OF ORGANISMS PER STATION, CLINCH RIVER, 1970

Station number	Early June	Late June	Early July	Late July	Early August	Late August
4	95			93		93
7	95	90	93	88	84	88
8	96	97	97	98	89	82
9	96	97	97	97	96	89
10	97	97	98	99	97	90
11	91			96		86

TABLE 29. ANALYTICAL TECHNIQUE, TYPE OF COMPARISON, AND THE NUMBER AND TYPE OF SIMILARITY COEFFICIENTS USED TO ANALYZE THE REDUCED 1970 CLINCH RIVER DATA SET

	Cluster analysis[a]		Nonmetric multidimensional scaling[b]	
	Q-mode	R-mode	Q-mode	R-mode
Unaffected stations	-	6	-	-
Affected stations	-	6	-	-
All stations	6	6	2	-
	Total dendrograms = 24		Total scatter diagrams = 2	

[a] Six coefficients were used: two correlation, two distance, and two presence-absence coefficients.

[b] Two coefficients were used: one distance and one presence-absence coefficient.

TABLE 30. COPHENETIC CORRELATION VALUES (r_{cc}) FOR 24 DENDROGRAMS COMPUTED FROM THE CLINCH RIVER DATA SET, JUNE TO AUGUST 1970

	S_J	S_{SM}	Corr 6	Corr 7	Dist 6	Dist 7
			R-mode			
Affected stations	0.97	0.84	0.73	0.63	0.76	0.97
Unaffected stations	0.97	0.91	0.63	0.63	0.69	0.92
All stations	0.95	0.76	0.75	0.75	0.78	0.97
			Q-mode			
All stations	0.93	0.77	0.82	0.70	0.91	0.80

TABLE 31. RESULTS OF Q-MODE CLUSTER ANALYSIS OF ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970: MINIMUM LEVEL OF SIMILARITY USED TO DEFINE CLUSTERS = 0.62; JACCARD'S COEFFICIENT

	Station						
Sampling Interval	4	7LB&MS	7RB	8	9	10	11
Cluster 1							
Before pH stress	X	X		X	X	X	
Immediately after pH stress		X					X
2 weeks after		X				X	
4 weeks after	X	X			X	X	X
6 weeks after				X	X	X	
8 weeks after	X			X	X	X	X
Cluster 2							
Before pH stress							
Immediately after pH stress				X	X	X	
2 weeks after				X	X		
4 weeks after				X			
6 weeks after							
8 weeks after							
Unclustered stations							
Before pH stress			X				
Immediately after pH stress			X				
2 weeks after			X				
4 weeks after			X				
6 weeks after		X	X				
8 weeks after		X	X				

TABLE 32. RESULTS OF Q-MODE CLUSTER ANALYSIS OF ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970: LEVEL OF SIMILARITY USED TO DEFINE CLUSTERS = 0.06; CORRELATION COEFFICIENT 6, $\sqrt{Y + 0.5}$ TRANSFORMATION AND STANDARDIZATION BY ROWS

Date	Station						
	4	7LB&MS	7RB	8	9	10	11
	Cluster 1						
Before pH stress	X	X	X				
Immediately after pH stress		X	X				
2 weeks after		X	X				
4 weeks after	X	X	X				
6 weeks after		X	X				
8 weeks after	X	X	X				
	Cluster 2						
Before pH stress				X			
Immediately after pH stress				X	X		
2 weeks after				X	X		
4 weeks after							
6 weeks after							
8 weeks after							
	Cluster 3						
Before pH stress							
Immediately after pH stress							
2 weeks after						X	
4 weeks after				X	X	X	X
6 weeks after				X	X	X	
8 weeks after				X	X	X	X
	Cluster 4						
Before pH stress					X	X	X
Immediately after pH stress						X	
2 weeks after							
4 weeks after							
6 weeks after							
8 weeks after							

TABLE 33. RESULTS OF Q-MODE CLUSTER ANALYSIS OF ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970: LEVEL OF SIMILARITY USED TO DEFINE CLUSTERS = 1.2; DISTANCE COEFFICIENT 6, $\sqrt{Y + 0.5}$ TRANSFORMATION AND STANDARDIZATION BY ROWS

	Station						
Date	4	7LB&MS	7RB	8	9	10	11
Cluster 1							
Before pH stress			X				
Immediately after pH stress			X	X	X	X	
2 weeks after			X	X	X		
4 weeks after	X		X				
6 weeks after			X	X	X		
8 weeks after	X	X	X	X	X		
Clusters 2 and 4 and unclustered stations (U)							
Before pH stress	U			U			
Immediately after pH stress		2					
2 weeks after		2				4	
4 weeks after		2		4	4	4	
6 weeks after		2					
8 weeks after		U					
Cluster 3							
Before pH stress					X	X	X
Immediately after pH stress							
2 weeks after							
4 weeks after							X
6 weeks after						X	
8 weeks after						X	X

TABLE 34. RESULTS OF Q-MODE CLUSTER ANALYSIS OF ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970: LEVEL OF SIMILARITY USED TO DEFINE CLUSTERS = 4.8; DISTANCE COEFFICIENT 7, $\sqrt{Y + 0.5}$ TRANSFORMATION AND STANDARDIZATION BY ROWS

	Station						
Date	4	7LB&MS	7RB	8	9	10	11
	Cluster 1						
Before pH stress	X	X		X	X	X	X
Immediately after pH stress							
2 weeks after							
4 weeks after	X						
6 weeks after						X	
8 weeks after	X					X	X
	Cluster 2						
Before pH stress			X				
Immediately after pH stress		X	X	X	X	X	
2 weeks after		X	X	X	X		
4 weeks after		X	X				
6 weeks after		X	X	X	X		
8 weeks after		X	X	X	X		
	Cluster 3 and unclustered stations						
Before pH stress							
Immediately after pH stress							
2 weeks after						X	
4 weeks after				X	X	U	U
6 weeks after							
8 weeks after							

TABLE 35. SUMMARY OF RESULTS OF Q-MODE CLUSTER ANALYSES, CUMBERLAND RIVER, 1973

Coefficients / Characteristics	S_J	S_{SM}	Dist 6	Dist 7	Corr 6	Corr 7
rcc	0.85	0.64	0.62	0.83	0.62	0.98
Similarity level of clusters[a]	1	3	1	2	1	3
Interpretable structure	good	good	poor	good	poor	fair
Highest associations	4 Oct 5 Oct	2 Jun 6 Jul	6 Jul 2 Aug	1 Oct 6 Oct	1 Oct 6 Oct	3 Jul 5 Jul
	3 Jun 6 Jun	3 Jun 6 Jun	2 Oct 6 Jul	2 Jul 2 Oct	6 Jul 2 Aug	1 Oct 6 Oct
	1 Jul 5 Jul	4 Oct 5 Oct	2 Jul 2 Oct 6 Jul 2 Aug	6 Jul 2 Oct	1 JUn 2 Jun	3 Jun 6 Jul
	1 Oct 6 Oct	1 Oct 6 Oct	2 Jul 2 Oct 6 Jul 2 Aug 3 Sep	3 Jul 5 Jul	4 Oct 5 Oct	4 Oct 5 Oct
	6 Aug 4 Oct 5 Oct	1 Jul 5 Jul	1 Oct 6 Oct	4 Aug 5 Oct	3 Jun 5 Aug	2 Aug 5 Aug
Principal causes of association (in order of importance) as indicated by analysis of biological[b] and chemical data	Station location and month sampled	Station location and month sampled	Station location	Station location and month sampled	Station location and month sampled	Month sampled and station location

TABLE 35 (continued)

Coefficients / Characteristics	S_J	S_{SM}	Dist 6	Dist 7	Corr 6	Corr 7
Principal causes of association as indicated solely by the highest biological associations[c]	Month sampled and station location	Month sampled and station location	Station location	Station location and month sampled	Month sampled and station location	Month sampled and station location

---

[a] 1 = distinct clusters form only at relatively low similarity levels;
2 = distinct clusters form at intermediate similarity levels;
3 = distinct clusters form at relatively high similarity levels.

[b] Results of cluster analysis were weighed heavily in this evaluation.

[c] Associations listed in characteristic four of this table.

TABLE 36. SUMMARY OF RESULTS OF Q-MODE CLUSTER ANALYSES, CUMBERLAND RIVER, 1975

Coefficients / Characteristics	S_J	S_{SM}	Dist 6	Dist 7	Corr 6	Corr 7
rcc	0.88	0.70	0.95	0.84	0.73	0.74
Similarity level of clusters[a]	1	2	2	2	1	2
Interpretable structure	good	good	poor	fair	poor	poor
Highest associations	3 Jun 3 Jul	3 Jun 3 Jul	3 Jun 3 Jul	3 Jun 3 Jul	3 Jun 3 Jul	3 Jun 3 Jul
	4 Jun 4 Jul	4 Jun 4 Jul	4 Jun 4 Jul	4 Jun 4 Jul	4 Jun 4 Jul	4 Jun 4 Jul
	5 Jun 5 Jul	5 Jun 5 Jul	5 Jun 5 Jul	5 Jun 5 Jul	5 Jun 5 Jul	5 Jun 5 Jul
	6 Jun 6 Jul	6 Jun 6 Jul	6 Jun 6 Jul	6 Jun 6 Jul	6 Jun 6 Jul	6 Jun 6 Jul
	7 Jun 7 Jul	7 Jun 7 Jul	7 Jun 7 Jul	7 Jun 7 Jul	7 Jun 7 Jul	7 Jun 7 Jul
	2 May 5 May	2 May 5 May	3 May 6 May	3 May 6 May	3 May 3 Aug	1 May 3 May

TABLE 36 (continued)

Coefficients Characteristics	S_{SM}	Dist 6	Dist 7	Corr 6	Corr 7	
Principal cause(s) of association (in order of importance) as indicated by analysis of both biological[b] and physicochemical data	Month in which collected	Month in which collected and station location	Month in which collected and station location	Month in which collected and station location	Month in which collected and station location	Month in which collected and station location
Principal cause(s) of association (in order of importance) as indicated solely by the highest biological associations[c]	Station location and month in which collected	Station location and month in which collected	Station location and month in which collected	Station location and month in which collected	Station location and month in which collected	Station location and month in which collected

[a] 1 = distinct clusters form only at relatively low similarity levels;
2 = intermediate;
3 = distinct clusters form at relatively high similarity levels.

[b] Results of cluster analysis were included in this evaluation.

[c] Associations listed in characteristic four of this table.

TABLE 37. RESULTS OF R-MODE CLUSTER ANALYSIS OF JACCARD'S (S_J) AND SIMPLE-MATCHING (S_{SM}) COEFFICIENTS, CLINCH RIVER DATA SET, 1970

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)				
Affected (S_J, 0.88)	Affected (S_{SM}, 0.88)	All stations (S_J, 0.77)	Affected (S_{SM}, 0.71)	Affected (S_J, 0.58)
Stenelmis Optioservus	Stenelmis Optioservus	Stenelmis Optioservus Dubiraphia Hydropsyche Chironomidae Corydalus cornutus Cheumatopsyche Acroneuria	Stenelmis Optioservus Microcylloepus Corydalus cornutus Chironomidae Dubiraphia Hemerodromia Hydropsyche Cheumatopsyche Promoresia Acroneuria	Stenelmis Optioservus Chironomidae Corydalus cornutus Microcylloepus Cheumatopsyche Dubiraphia Hemerodromia Hydropsyche Promoresia Acroneuria Goniobasis spinella
Chironomidae Corydalus cornutus	Microcylloepus Corydalus cornutus Chironomidae	Microcylloepus Promoresia Hemerodromia	Isonychia Baetis	Isonychia Baetis Tricorythodes Hetaerina
Microcylloepus Cheumatopsyche Dubiraphia Hemerodromia	Dubiraphia Hemerodromia Hydropsyche Cheumatopsyche	Isonychia Baetis	Ephemerella Goniobasis carinifera Tricorythodes Hetaerina	Ephemerella Goniobasis carinifera
	Tricorythodes Hetaerina	Ephemerella Goniobasis carinifera	Simulium Anculosa	
		Tricorythodes Hetaerina		

TABLE 38. RESULTS OF R-MODE CLUSTER ANALYSIS OF DISTANCE COEFFICIENTS, CLINCH RIVER DATA SET, 1970

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)			
Affected (Dist 7, 1.8)	Affected (Dist 7, 5.3)	All stations (Dist 7, 4.2)	Unaffected (Dist 7, 2.3)
Promoresia	*Stenelmis*	*Microcylloepus*	*Microcylloepus*
Hemerodromia	*Optioservus*	*Corydalus cornutus*	*Hemerodromia*
		Chironomidae	*Hetaerina*
		Baetis	*Promoresia*
			Pleuroceridae
			Sphaerium
			Psephenus herricki
			Anculosa subglobosa
			Goniobasis spinella
			Acroneuria
Psephenus herricki	*Microcylloepus*	*Promoresia*	*Ephemerella*
Potamanthus	*Corydalus cornutus*	*Hemerodromia*	*Corydalus cornutus*
Anculosa subglobosa	*Baetis*	*Psephenus herricki*	
Goniobasis carinifera	Chironomidae	*Anculosa subglobosa*	
Anculosa	*Dubiraphia*	Pleuroceridae	
Goniobasis spinella	*Promoresia*	*Sphaerium*	
Stenonema	*Hemerodromia*	*Goniobasis spinella*	
Perlesta placida	*Psephenus herricki*	*Heptagenia*	
Ephemerella	*Potamanthus*	*Hetaerina*	
Tricorythodes	*Anculosa subglobosa*	*Acroneuria*	
Hetaerina	*Goniobasis carinifera*	*Perlesta placida*	
	Anculosa	*Ephemerella*	
	Goniobasis spinella	*Tricorythodes*	
	Stenonema	*Ephoron*	

TABLE 38 (continued)

Affected (Dist 7, 1.8)	Affected (Dist 7, 5.3)	All stations (Dist 7, 4.2)	Unaffected (Dist 7, 2.3)
	Perlesta placida	*Potamanthus*	
	Ephemerella		
	Tricorythodes		
	Hetaerina	*Stenonema*	
	Acroneuria	*Anculosa*	
		Goniobasis carinifera	

TABLE 39. RESULTS OF R-MODE CLUSTER ANALYSIS OF S_{SM} AND S_J, CLINCH RIVER STATIONS UNAFFECTED BY THE 1970 pH STRESS

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)	
Unaffected (S_{SM}, 0.89)	Unaffected (S_J, 0.87)
Stenelmis *Optioservus* *Dubiraphia* *Ephemerella* *Baetis* *Hydropsyche* *Psephenus herricki* *Goniobasis spinella*	*Stenelmis* *Optioservus* *Dubiraphia* *Ephemerella* *Baetis* *Hydropsyche* *Psephenus herricki* *Goniobasis spinella*
Stenonema *Isonychia* *Goniobasis carinifera* *Acroneuria* *Tricorythodes*	*Stenonema* *Isonychia* *Goniobasis carinifera* *Acroneuria* *Tricorythodes*
Chironomidae *Corydalus cornutus* *Cheumatopsyche* *Potamanthus*	Chironomidae *Corydalus cornutus* *Cheumatopsyche* *Potamanthus*
Anculosa *Anculosa subglobosa*	*Anculosa* *Anculosa subglobosa*

TABLE 40. TROPHIC-FUNCTIONAL CODES FOR TAXA CLUSTERED BY R-MODE CLUSTER ANALYSIS OF JACCARD'S (S_J) AND SIMPLE-MATCHING (S_{SM}) COEFFICIENTS, CLINCH RIVER DATA SET, 1970

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)				
Affected (S_J, 0.88)	Affected (S_{SM}, 0.88)	All stations (S_J, 0.77)	Affected (S_{SM}, 0.71)	Affected (S_J, 0.58)
3524 4613	3411 3524 4613	2224 3411 3524 4613 4624	2224 3411 3524 4613 4624 4713	2224 3411 3524 4613 4624 4713
2224 3411 4713	2224 3411 4713	3411 4713		
		2224 3424	2224 3424	2224 2312 3424 4613
		3411 3424	2312 3411 3424 4613	3411 3424
	2312 3424	2313 4613	2212 3411	

TABLE 41. TROPHIC-FUNCTIONAL CODES FOR TAXA CLUSTERED BY R-MODE CLUSTER ANALYSIS OF DIST 7 COEFFICIENTS CLINCH RIVER DATA SET, 1970

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)			
Affected (Dist 7, 1.8)	Affected (Dist 7, 5.3)	All stations (Dist 7, 4.2)	Unaffected (Dist 7, 2.3)
3411 4713	3411	3411 3424 3524 4613	2212 3411 3412 4613 4624 4713
2312 2612 3411 3412 3424 4613 4624	2312 2612 3411 3412 3424 3524 4613 4624 4713	2212 2312 2612 3411 3412 3424 4613 4624 4713	3424 4613
		3411 3424	

TABLE 42. TROPHIC-FUNCTIONAL CODES FOR TAXA CLUSTERED BY R-MODE CLUSTER ANALYSIS OF SIMPLE-MATCHING (S_{SM}) AND JACCARD'S (S_J) COEFFICIENTS, CLINCH RIVER, 1970

Unaffected[a] (S_J[b], 0.87[c] and S_{SM}[b], 0.89[c])
2224
3411
3412
3424
2224
2312
3411
3424
4624
2224
2612
3524
4613
3411

---

[a] Stations unaffected by spill.

[b] Clustering coefficient utilized.

[c] Similarity level at which clusters formed.

TABLE 43. RESULTS OF R-MODE CLUSTER ANALYSIS OF S_J AND S_{SM} AFTER REORDERING CLUSTERS ACCORDING TO TROPHIC-FUNCTIONAL CODES, CLINCH RIVER, 1970

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)				
Affected (S_J, 0.88)	Affected (S_{SM}, 0.88)	All stations (S_J, 0.77)	Affected (S_{SM}, 0.71)	Affected (S_J, 0.58)
Optioservus	*Optioservus*			*Cheumatopsyche*
Stenelmis	*Stenelmis*		*Cheumatopsyche*	*Hydropsyche*
			Hydropsyche	*Dubiraphia*
		Cheumatopsyche	*Dubiraphia*	*Microcylloepus*
		Hydropsyche	*Microcylloepus*	*Optioservus*
		Dubiraphia	*Optioservus*	*Promoresia*
		Optioservus	*Promoresia*	*Stenelmis*
	Microcylloepus	*Stenelmis*	*Stenelmis*	*Goniobasis spinella*
Chironomidae	Chironomidae	Chironomidae	Chironomidae	Chironomidae
Corydalus cornutus	*Corydalus cornutus*	*Corydalus cornutus*	*Corydalus cornutus*	*Corydalus cornutus*
		Acroneuria	*Acroneuria*	*Acroneuria*
			Hemerodromia	*Hemerodromia*
Cheumatopsyche	*Cheumatopsyche*			
Hydropsyche	*Hydropsyche*	*Microcylloepus*		
Dubiraphia	*Dubiraphia*	*Promoresia*		
Microcylloepus	*Hemerodromia*	*Hemerodromia*		
Hemerodromia				
		Isonychia	*Isonychia*	*Isonychia*
		Baetis	*Baetis*	*Tricorythodes*
				Baetis
				Hetaerina

TABLE 43 (continued)

Affected (S_J, 0.88)	Affected (S_{SM}, 0.88)	All stations (S_J, 0.77)	Affected (S_{SM}, 0.71)	Affected (S_J, 0.58)
		Goniobasis carinifera *Ephemerella*	*Tricorythodes* *Goniobasis carinifera* *Ephemerella* *Hetaerina*	*Goniobasis carinifera* *Ephemerella*
	Tricorythodes *Hetaerina*	*Tricorythodes* *Hetaerina*		
			Simulium *Anculosa*	

TABLE 44. RESULTS OF R-MODE CLUSTER ANALYSIS OF DIST 7 COEFFICIENTS AFTER REORDERING CLUSTERS ACCORDING TO TROPHIC-FUNCTIONAL CODES, CLINCH RIVER, 1970

CATEGORY OF STATIONS INCLUDED (COEFFICIENT, SIMILARITY LEVEL AT WHICH CLUSTERS FORMED)

Affected (Dist 7, 1.8)	Affected (Dist 7, 5.3)	All stations (Dist 7, 4.2)	Unaffected (Dist 7, 2.3)
Promoresia	*Optioservus*	*Microcylloepus*	
Hemerodromia	*Stenelmis*	*Baetis*	
		Chironomidae	
	Tricorythodes	*Corydalus cornutus*	
	Potamanthus		
	Dubiraphia	*Sphaerium*	*Sphaerium*
Tricorythodes	*Microcylloepus*	*Tricorythodes*	*Microcylloepus*
Potamanthus	*Promoresia*	*Ephoron*	*Promoresia*
Anculosa	*Anculosa*	*Potamanthus*	*Anculosa subglobosa*
Anculosa subglobosa	*Anculosa subglobosa*	*Promoresia*	*Goniobasis spinella*
Goniobasis carinifera	*Goniobasis carinifera*	*Anculosa subglobosa*	Pleuroceridae
Goniobasis spinella	*Goniobasis spinella*	*Goniobasis spinella*	*Psephenus herricki*
Psephenus herricki	*Psephenus herricki*	Pleuroceridae	*Hetaerina*
Ephemerella	*Baetis*	*Psephenus herricki*	*Acroneuria*
Stenonema	*Ephemerella*	*Ephemerella*	*Hemerodromia*
Hetaerina	*Stenonema*	*Heptagenia*	
Perlesta placida	Chironomidae	*Hetaerina*	
	Corydalus cornutus	*Acroneuria*	
	Hetaerina	*Perlesta placida*	
	Acroneuria	*Hemerodromia*	
	Perlesta placida		
	Hemerodromia		
			Ephemerella
			Corydalus cornutus

TABLE 45. CLUSTERS OF TAXA DEFINED BY R-MODE CLUSTER ANALYSIS OF JACCARD'S AND THE SIMPLE-MATCHING COEFFICIENTS, CLINCH RIVER, 1970

Unaffected stations	Affected stations
Stenelmis	*Stenelmis*
Optioservus	*Optioservus*
Dubiraphia	
Ephemerella	*Hydropsyche*
Baetis	*Cheumatopsyche*
Hydropsyche	*Dubiraphia*
Psephenus herricki	
Goniobasis spinella	*Isonychia*
	Baetis
Stenonema	
Isonychia	*Goniobasis carinifera*
Acroneuria	*Ephemerella*
Tricorythodes	
Chironomidae	Chironomidae
Corydalus cornutus	*Corydalus cornutus*
Cheumatopsyche	
Potamanthus	*Tricorythodes*
	Hetaerina
Anculosa	
Anculosa subglobosa	

SECTION 6

ORDINATION--NONMETRIC MULTIDIMENSIONAL SCALING

6.1 GENERAL DESCRIPTION

Ordination is an analytical technique often used when dealing with large sets of data. In aquatic ecology, ordination involves plotting either the samples or the species found in the samples in a two- or three-dimensional diagram. The choice of axes on which the points are plotted depends on the method of ordination selected. In most ordinations, either samples are plotted in a space where the axes are species or species are plotted in a space defined by the samples. The primary advantage of ordination over cluster analysis is that it "allows one to examine a scatter diagram displaying a summary of the structure of the data without having to first assume that clusters are present" (Rohlf, 1970). In this sense, ordination is a powerful alternative to cluster analysis.

Several techniques of ordination are available, including principal component ordination, principal coordinate ordination, the Bray-Curtis polar ordination technique (Whittaker, 1975), and nonmetric multi-dimensional scaling (MDS) (Kruskal, 1964*a* and 1964*b*). Only nonmetric MDS is dealt with in this study because it seems ideally suited to the kinds of data obtained from biological surveys. Specifically, nonmetric MDS is computationally robust when data are missing; it can accommodate quantitative, ranked, or presence-absence data; and it can be used with any kind of measure of correlation, similarity, or distance.

6.2 ANALYTICAL PROCEDURES

In nonmetric MDS, similarities or distances between samples are treated on the *ordinal scale*; that is, they are ordered from smallest to largest. A configuration is then found in which "rank order of (ratio scaled) distances best produces the original input *ranks*. One tries to do this in the lowest dimensionality that produces a 'close enough' ordinal fit" (Green and Carmone, 1970). The important point is that the MDS program operates on ranked similarities and distances rather than on actual similarities. Thus, nonmetric MDS is perfectly applicable to matrices of similarity and distance computed from ranked data or even presence-absence data.

For a specified number of dimensions chosen in advance by the investigator, the computer programs "try to find a configuration of points whose interpoint distances are monotone--that is, have the same (or possibly the inverse) ranks as the input data" (Green and Carmone, 1970). The coordinates of this new configuration are the values used in the ordination. In practice, perfect configurations are unusual. The measure of departure from monotonicity is called *stress*. The higher the stress, the less nearly perfect the degree of monotonicity. In general, the stress decreases with the number of dimensions and the number of iterations used in the analysis.

6.3 RESULTS

6.3.1 Q-Mode Analysis--Clinch River Data

6.3.1.1 Presence-Absence Data--

Two applications of nonmetric MDS, presence-absence and quantitative data, were tested (as an alternative to cluster analysis). Figure 36 shows the Q-mode ordination of presence-absence data from 36 zoomacrobenthic samples from the Clinch River. In this three-dimensional display, one large, fairly tightly grouped set of 16 samples lies near the right-center portion of the diagram, indicating the faunal similarity of these samples. This group of samples is quite similar to the first cluster in Table 31, which shows the results of cluster analysis of the matrix of Jaccard's coefficients (S_J). Samples outside the main group in Figure 36 are from station 8, six and eight weeks after the spill; station 9, four, six, and eight weeks after the spill; and station 10, two weeks after the spill.

Samples collected at stations 8, 9, and 10 after the spill tended to form individual groupings near the upper center of Figure 36. For example, samples collected at station 9 four, six, and eight weeks after the spill grouped together. Since the biological communities at these stations were exposed to different pH conditions for varying lengths of time, they were expected to stand alone, not to cluster together.

Samples from substation 7RB did not form any distinct pattern, but were scattered along (left side of Figure 36). This agreed with the results obtained from Q-mode clustering, where samples from 7RB did not cluster at the 0.62 level of similarity (reference Table 31). Samples from 7LB&MS, collected six and eight weeks after the spill, were also widely separated. The lack of similarity of these samples with the other 34 samples also agreed with the Q-mode clustering results presented in Table 31.

6.3.1.2 Quantitative Data (Species Counts)--

When species counts were used in nonmetric MDS, the 36 zoomacrobenthic samples appeared to be randomly scattered (Figure 37). Only one recognizable group of stations was present. This cluster, near the upper left margin of the diagram, consisted primarily of samples from substation 7RB. Two samples from 7LB&MS, collected six and eight weeks after the spill, were also in this cluster, indicating that the left bank and midstream sections of station 7 were similar to the right bank.

The results of cluster analysis of the distance coefficient Dist 7 (Table 34) were compared with Figure 37 to determine any similarities between clustering and the ordination of quantitative data. At a level of similarity of 4.8, three clusters were formed in the Dist 7 dendrogram; (1) 11 samples from either the control stations (stations 4 and 11) or stations 8, 9, and 10 before the spill; (2) 20 samples from stations 7 through 10, including station 7LB&MS; and (3) samples from stations 8 and 9, four weeks after the spill, and station 10, two weeks after the spill. Figure 37 has only one poorly defined group of nine stations: station 8, immediately after the spill and two and eight weeks after the spill; station 9, immediately after the spill and two, six, and eight weeks after the spill; station 7LB&MS, immediately after the spill and two weeks after the spill; and station 11, six weeks after the spill. It could be equated with the second cluster in Table 34, but only in a

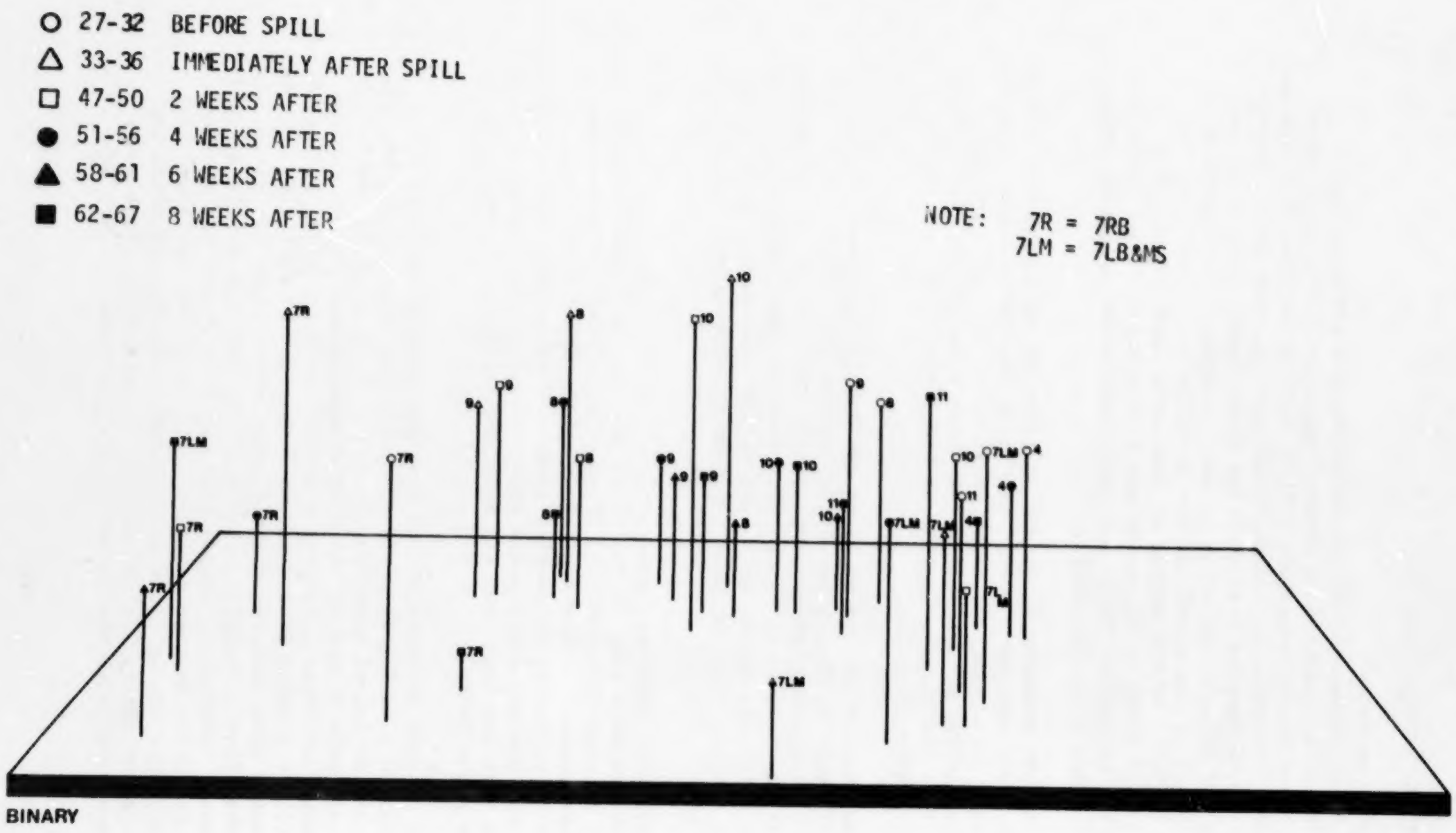


Figure 36. Three-dimensional ordination by nonmetric multidimensional scaling computed from distance coefficients based on presence-absence data, showing faunal similarities between samples collected from the Clinch River in 1970.

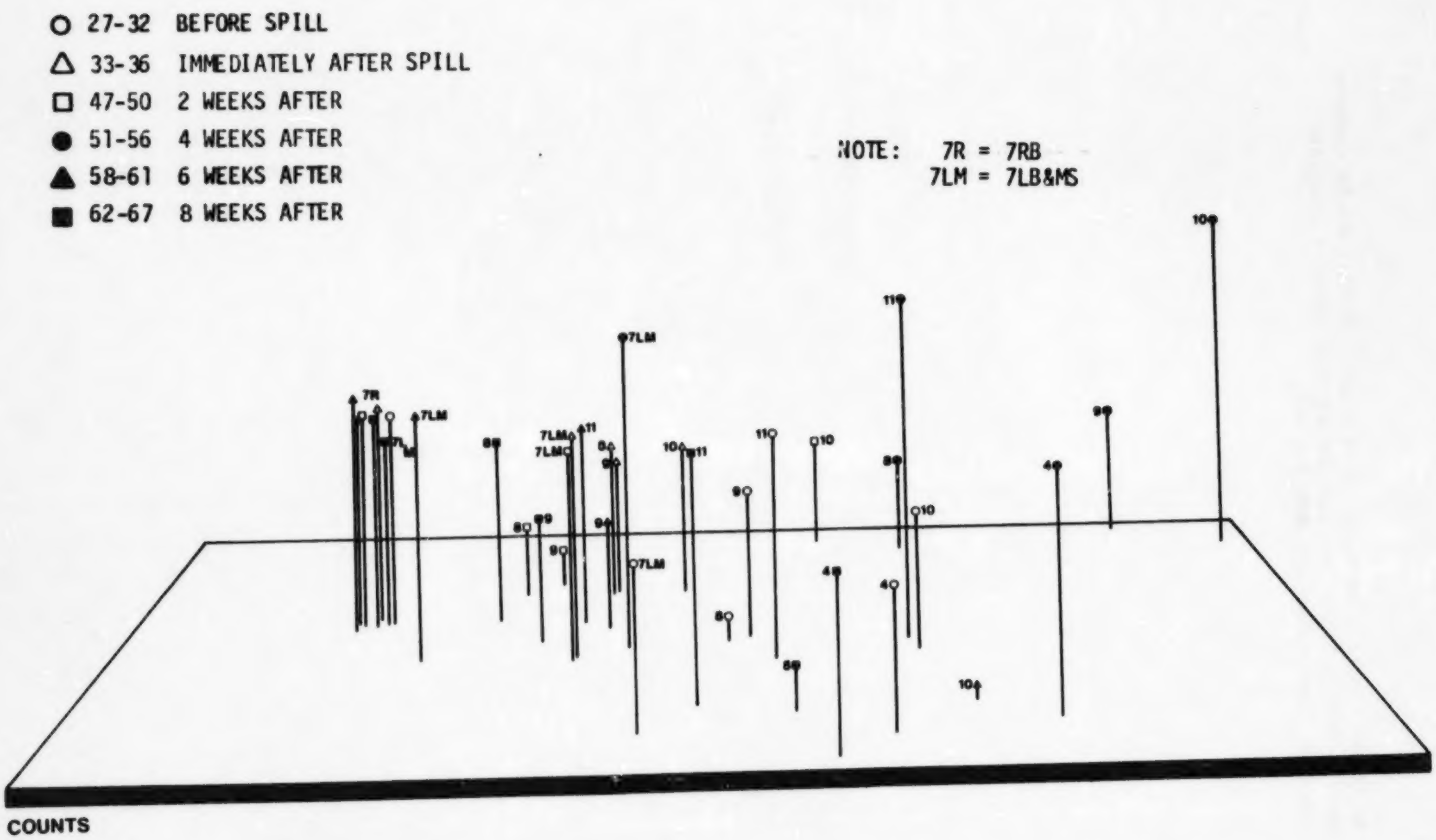


Figure 37. Three-dimensional ordination of nonmetric multidimensional scaling computed from distance coefficients based on counts of species, showing faunal similarities between samples collected from the Clinch River in 1970.

general sense. No other clusters were found in Figure 37 that would compare favorably with those noted in Table 34, indicating that cluster analysis tended to group stations that nonmetric MDS did not.

SECTION 7

DIVERSITY INDICES

7.1 GENERAL DESCRIPTION

The third kind of analysis of community structure considered was the use of indices of species diversity, and indices of diversity at higher taxonomic levels. A diversity index is a statistic that combines information on both the number of species in an assemblage and the evenness of the distribution of individuals among those species (Pielou, 1969). A high index of species diversity results from the presence of many species with nearly even abundances; a low index results from a few species in an assemblage dominated by one species. A diversity index of intermediate value can result either from a few species with nearly even distributions or from many species with uneven distributions. Therein lies the method's greatest weakness: that a value of species diversity can result from study of assemblages with quite different distributions of species. A second weakness is that, unlike cluster analysis and ordination, which operate on similarity coefficients, diversity indices are not affected by the species present, only by their numbers. Thus, although quite different communities can be compared, equal diversity does not imply equal tolerance to potential environmental impacts.

Failure to recognize the importance of these two inherent weaknesses has led to widespread misuse of diversity indices and misunderstanding of the concept of diversity. As a result, an extensive literature exists on the disadvantages of their use and the drawbracks of the concept (see e.g., Hamilton, 1975; Hedgpeth, 1973). Diversity indices remain, however, as one tool available for ecologists. Like any tool they can be misused, and getting the most out of them requires an understanding of their limitations. Like many other tools, they are most valuable when used in conjunction with other information or tools. The concept of diversity has been widely applied in all branches of ecology (Woodwell, 1970), and its use for evaluating potentially impacted communities in applied aquatic ecology is particularly noteworthy. Former EPA administrator R. E. Train (1973) has stated that "for top management and general public policy development, monitoring data must be shaped into easy-to-understand indices that aggregate data into understandable forms. Failure to do so will result in suboptimun achievement of goals at much greater expense." Whitten (1975) has concluded that "the most suitable means of analyzing community structure for the purpose of pollution assessment appears to be the diversity index."

Although most of the theory of diversity of communities has been based on species diversity, applied ecologists often have to enumerate organisms at higher taxonomic levels because of the difficulty of identifying species (see earlier discussions, section 5.3.2.1) and, consequently, the high cost of doing so. Results of such analyses, while internally consistent within a study, are by no means comparable from study to study. Therefore, the search for absolute values of diversity are unrealistic and inconsistent with the purpose of applied aquatic ecology. Nonetheless, in the examples that follow, we continue to refer to species diversity as a concept although the diversity values are in fact sometimes based on undifferentiated higher taxa, especially for such groups as oligochaetes and chironomids.

7.2 ANALYTICAL PROCEDURES

Measures of species diversity can be grouped into four general types: (1) species richness; (2) indices that assume a particular distribution, for example Williams' α index (Fisher et al., 1943); (3) indices measuring probability of encounter, such as Simpson's (1949) index and the sequential comparison index (SCI) (Cairns et al., 1968); and (4) indices adapted from information theory, Shannon's index (Shannon and Weaver, 1949), Brillouin's index (Brillouin, 1962), and the approximate index (e.g., Wilhm and Dorris, 1968).

Species richness (S) is simply the number of species present in a sample or community. This measure of diversity has the same drawbacks as the use of presence-absence data in cluster analysis and ordination. It makes no distinction, for example, between a sample with 1 individual each of species A and B, a sample with 1 individual of species A and 1000 of species B, and a sample with 1000 of each species. Moreover, it is highly dependent on sample size.

Dennison and Hay (1967) presented an equation that allows one to compute the number of individuals one must collect to be certain (with a given probability of error) of collecting species that comprise a given proportion of the community.

$$N = \frac{\log (E)}{\log (1 - p)}$$

where E = the probability of error the investigator is willing to accept, p = the proportion each target species comprises of the community, and N = the number of individuals to be collected. For example, to be 95 percent certain (i.e., 5 percent error: E = 0.05) of collecting species that comprise 1 percent of the community (p = 0.01),

$$N = \frac{\log (0.05)}{\log (1 - 0.01)} = \frac{-1.30}{-0.00436} = 298 \text{ individuals.}$$

The use of this equation shows that rather large samples are needed to assure that even relatively common species are not missed by chance alone and casts doubt on the value of species richness as a measure of species diversity in applied aquatic ecology.

Williams' α index (Fisher et al., 1943) is based on the assumption that the abundances of organisms in the community being studied fit a logarithmic series. Williams (1950) proposed that the parameter of such distributions be used as an index of diversity of the community. The parameter α is an intrinsic property of communities, unaffected by sample size, that is proportional to the number of species present.

$$S = \alpha \log_e (1 + \frac{N}{a})$$

where α = the parameter of the logarithmic series used to measure species diversity, N = the number of individuals in the sample, and S = the number of species in the sample. Thus, although α is a property of the community that is independent of sample size, since its estimation is based on number of individuals and species in a sample, the size of the sample is of paramount importance. This index can be applied only if species abundance fits a logarithmic series, which may be impossible to determine when only a few species are present. Thus, as Pielou (1969) has pointed out, "α is unsuitable as a diversity index unless the collection at hand has many species and also unless abundances form a logarithmic series."

Simpson's (1949) index was derived from probability theory and answers the question "What is the probability that two specimens picked at random from a community of infinite size are the same species?" If a species i is present in the community in the proportion p_i, the probability of selecting two individuals of species i at random is the joint probability of p_i^2. Simpson's equation is:

$$D = 1 - \sum_{i=1}^{S} (p_i^2)$$

where S = the number of species, p_i = the proportion of individuals in the ith species, and D = the diversity. As with some of the other indices discussed below, the correct computation of Simpson's diversity requires a fully censused community, which never occurs in practical ecology. Moreover, Krebs (1972) has noted that Simpson's index gives relatively little weight to rare species, a characteristic that is undesirable where rare species have great impact on the community, as is the case with predators.

A problem universally associated with the use of diversity indices has been the time and level of professional expertise required for taxonomic identification of organisms in the sample. Cairns et al. (1968) developed the sequential comparison index (SCI) to overcome this problem. Given a random sample of specimens, $A_1, A_2, \ldots A_N$, sequential comparisons are made between organisms (e.g., A_1 vs. A_2, A_2 vs. A_3), and the number of runs or consecutive specimens of the same species is determined. The ratio of number of runs divided by (N + 1), where N = the number of organisms, is the measure of diversity. Patil and Taillie (1976) have demonstrated that, with a minor correction for bias, the SCE becomes an unbiased estimator of Simpson's index diverted from probability theory.

Margalef (1956) proposed analysis of mixed-species communities by methods derived from information theory. In this sense, diversity is equated with the uncertainty that exists when a single organism is selected at random from the community. The more species present in a community and the more even their distribution, the greater the uncertainty and the larger the species diversity. Information content is a measure of uncertainty and is a reasonable measure of diversity. The most commonly employed diversity indices derived from information theory are Shannon's index (H'), Brillouin's index (H), and the approximate index (H"). Pielou (1966, 1967, 1969, 1974, 1975,

1977) has elucidated the theory behind the use of these indices in ecology, and Kaesler and Herricks (1977) and Kaesler et al. (1978), drawing on her work, have evaluated their use in applied aquatic ecology.

Shannon and Weaver (1949) introduced the following equation for the information content per symbol of a code made up of S different symbols, each with a probability of occurrence of p_i:

$$H' = - \sum_{i=1}^{S} (p_i \log_e p_i).$$

In an ecological context, S = the number of species in a conceptually infinite community and p_i = the proportion the ith species comprises of the community. Note that this equation is not intended for use with sample data but rather requires knowledge of the proportions each species comprises of the community, information that is never available to the applied aquatic ecologist. In their evaluation of this index of diversity, Kaesler and Herricks (1977) wrote: "The primary problem with the use of this equation is that in the real world it is usually impossible to define or to sample randomly from the conceptually infinite population. How, for example, would one define the limits of a community occupying the polluted reaches of a stream in which downstream recovery and accommodation, seasonal change, and recruitment due to stream drift are important factors? Moreover, the probabilities of occurrence or the proportion of each species in the community can never be known, and a reasonably precise estimation may require unreasonably large samples, especially for rare species."

Brillouin's equation gives the information content per symbol of a message and is thus based on samples rather than conceptually infinite communities:

$$H = \frac{1}{N} \log_e \frac{N!}{N_1!N_2! \ldots N_S!}$$

In an ecological context, S = the species in a sample, N = the number of organisms in the sample, and N_i = the number of organisms belonging to the ith species. Kaesler and Herricks (1977) wrote: "Brillouin's H is the species diversity per individual of a collection in which all N specimens have been assigned to one of s species and counted to give the N_i's. This equation has not been popular, partly because the factorials involved often become astronomically large, but with readily available, high-speed computers, and tables of logarithms of factorials it need no longer be avoided. Moreover, it is important because it, and not Shannon's equation, gives the actual diversity of a fully censused collection of organisms. It is not a statistical estimate but an actualy measurement of the diversity of the working ecologists basic unit--the sample."

The approximate index is given by:

$$H'' = - \sum_{i=1}^{S} \frac{N_i}{N} \log_e \frac{N_i}{N}$$

where S = the number of species in a sample and N_i/N = the proportion of the ith species in the sample. Kaesler and Herricks (1977) have pointed out that the approximate index "is the one that is used most often in practice. It resembles Shannon's equations, but the probabilities of occurrence of species have been replaced by their proportion in the sample, N_i/N. In spite of its popularity, the approximate index has some serious drawbacks that have not always been thoroughly appreciated by some of its users.

"Pielou (1966) has shown that this equation can be used for two quite different purposes. Unfortunately, it suits neither purpose particularly well. First, it has been used to estimate Shannon's diversity, H', when the N_i's are sample values. H" is, in fact, a maximum likelihood estimator of H', but it is also a biased estimator. An appropriate correction term has been derived, but in order to apply it, it is necessary to know s, the number of species in the community (in the universe, not the sample) (Basharin, 1959). It is, of course, not possible to know the number of species in a real community without making a complete census of the community, which, even if it were possible, would not be consistent with the purposes of environmental protection. Wilhm (1968) has pointed out that for a finite s the bias approaches zero as the number of individuals in the sample, N, approaches infinity; and Peet (1974) has suggested that the bias is small for most ecological applications. However, when equal-effort sampling methods are used, one is not assured of large samples, particularly from heavily polluted stations. Thus, the very samples in which the applied ecologist is most interested will in many instances be those in which the bias is most severe.

"Second, H" has been used as a substitute for Brillouin's H to approximate (not estimate in a statistical sense) the reputedly cumbersome H in order to avoid computing fractorials. This is done when the N_i's are regarded as population values. The equation for H"...can be derived from Brillouin's equation,...but the derivation is valid only when all the N_i's are...very large indeed...(Pielou, 1969). The reason for this stipulation of very large N_i's is that the derivation of [H"]...from Brillouin's equation...depends on the substitution of N(ln N-1) for lnN! in...[H]. For small values of N, this substitution is simply not warranted. Especially with equal-effort sampling, one cannot be sure that all the N_i's will be very large, and again the low diversity samples that are usually of greatest interest to applied ecologists will be incorrect--in this case with artifically high values because H" is always larger than H.

"Wilhm (1968) has argued that H seldom applies to ecological measures of diversity because the N_i's "...are rarely population values and must be estimated from the sample." But, as was pointed out earlier, this is entirely a matter of defining the population (universe). If ecological samples are regarded as fully-censused collections, then H ... gives the exact diversity not an estimate or approximate. Because it is usually not possible to define the limits of a population, especially in stream surveys, nor to sample at random from it, we judge that it is most appropriate to regard a sample as a message from the ecosystem. That message has an information content or diversity that is appropriately computed with Brillouin's equation..."

Pielou (1967) has shown that diversity indices from information theory may be studied at each level in the taxonomic hierarchy and that the components at various levels are independent and additive. Thus, for example, one may study diversity of orders, families within orders, genera within families, and species within genera. Furthermore, one may devise other schemes of classification that provide greater ecological insight than the taxonomic heirarchy. In this study two nontaxonomic hierarchical classifications were used in addition to the taxonomic hierarchy. The first, adapted from Cummins' (1973) work, emphasized the trophic-functional (TF) role of each organism in the stream ecosystem. The major categories used in the classification are functional group, feeding mechanism, dependence, and principal food habit (Table 11). To adapt this scheme to our needs, a numerical code was assigned to each designation in a category and a hierarchy established (Table 12). The second classification scheme was based on the observations of Steinmann (1907, 1908) and Hynes (1960, 1970), who related anatomical and behavioral adaptations of benthic invertebrates to habitat preference and the organism's ability to obtain food. This system used three categories that stressed functional morphology: head position, H; general body shape, B, and type and shape of respiratory organs, R (Table 13).

7.3 DIVERSITY OF SAMPLES FROM THE CLINCH RIVER

Species diversity of 36 samples collected in 1970 from the Clinch River were computed using Brillouin's H and, for comparison, the approximate index H".

7.3.1 Species Diversity

Results of diversity analysis with Brillouin's index (H) are summarized in Figure 38, in which time in days is plotted along the abscissa and stations along the ordinate. The ordinate was complicated somewhat by dividing station 7 into substations 7RB (right bank) and 7LB&MS (left bank and midstream). Both substations were at the same river kilometer, but 7LB&MS was plotted closer to station 4 because it was unaffected by the spill and day-to-day operation of the power plant. Substation 7RB, on the other hand, was strongly affected by both because the effluent from the power plant flowed along the right bank.

H was rounded to one decimal place, values were plotted, and diversity was contoured at an interval of 0.4 to illustrate geographic and temporal differences between stations or groups of stations. The most discernible differences involved the periods before and after the spill and substations 7LB&MS and 7RB (Figure 38). The contours of diversity near these samples are closely spaced, indicating a pronounced change in community structure.

After the spill, diversity decreased to between 1.6 and 1.8 at the affected stations. This decrease was followed by an increase in diversity over the next 60 days until H equaled or approximated the values found before the pH stress. Examination of the samples contained within each contour line indicated a sequence of recovery for the affected stations. Biological recovery through time began at the farthest downstream station and proceeding upstream. This is contrary to the sequence postulated by some investigators, who claim that biological recovery begins at the site closest to healthy sites. Although their hypothesis may explain the recovery of some streams, it does not in this instance. In this study, biological recovery depended more on the severity of the initial damage than on a site's proximity to unaffected tributaries or headwater areas.

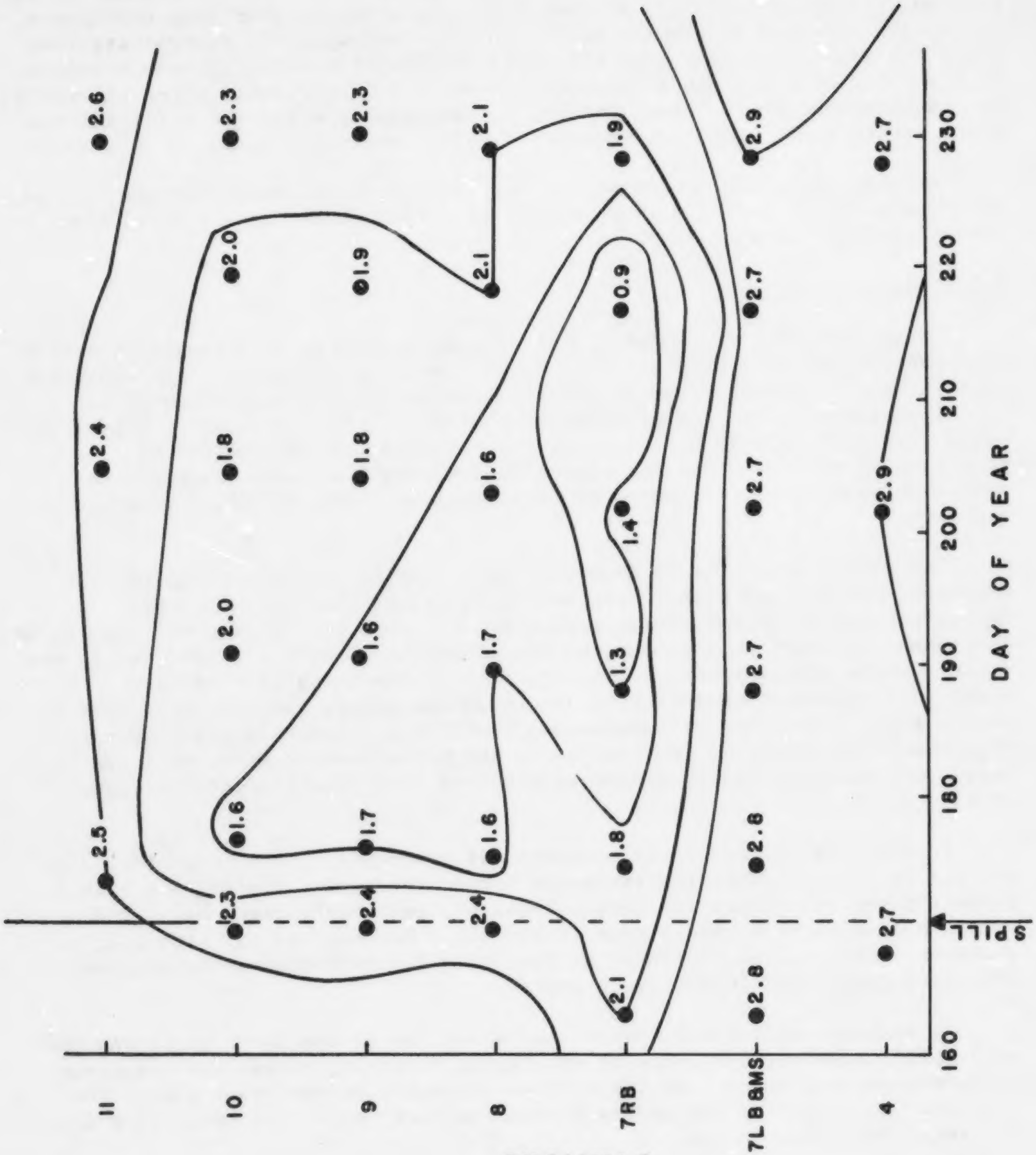


Figure 38. Species diversity (Brillouin's H) of 36 zoomacrobenthic samples from the Clinch River, 1970; contour interval 0.4.

Values of H tabulated in Table 46 form groups that are remarkably similar to those obtained from clustering presence-absence data with Jaccard's coefficient (Table 31). In both analyses, the upstream and downstream control stations (stations 4 and 11) are similar. Moreover, samples from stations 8, 9, and 10 collected before the pH stress are similar to the control stations. After the spill, the most severely impacted sites tended to cluster together, as evidenced by the cluster containing station 8, immediately after the spill and two and four weeks later; station 9, immediately after the spill and two weeks later; and station 10, immediately after the spill.

Results using the approximate index H" (Figure 39, Table 47) are similar, but H" appears to depend more on sample size than other investigators have thought (Wilhm and Dorris, 1968).

7.3.2 Hierarchical Diversity

It is important to remember that species diversity is a single statistic that cannot provide a complete picture of community structure. The statistic does, however, provide a useful quantitative measure of environmental conditions, and it has characteristics that allow it to be used as a heuristic tool. Pielou (1967, 1969, 1975) has demonstrated the possibility of partitioning species diversity into additive components that express the amount of diversity contributed by each level or component of a taxonomic hierarchy.

The use of hierarchical diversity as a heuristic tool enables one to determine whether additional insight can be gained from species-level determinations in biomonitoring programs. For example, if only one species of each genus is found in a community, the component of diversity due to species within genera equals zero. Similarly, if only one genus of each family is found, the component of diversity due to genera within families also equals zero. Thus, under some circumstances, costly and time-consuming species determinations may be of limited value, and environmental assessment can be based on discrimination of higher taxa rather than identification of lower ones.

To test the usefulness of hierarchical diversity, the Clinch River data set was partitioned into the taxonomic categories order, family, and genus. Diversity was calculated for each substation, left bank, right bank, and midchannel sections of the stream (Table 48). For the control stations, stations 4 and 11, the component of diversity for each taxonomic category decreased from order to family to genus.

At stations affected by day-to-day operation of the power plant and the acid spill, however, the trend of decreasing diversity within the taxonomic hierarchy was not found. At the stations impacted by the power plant, the component of diversity for genera within families contributed more to H than did families within orders.

In previous studies of the Clinch River, Crossman et al. (1973) noted that station 8 was chronically stressed by the power plant and was the most severely impacted station after the pH stress. Stress not only caused a reduction in the number of taxa, but it changed the relative abundance of each species. As a result, diversity decreased. At substation 8LB the diversity

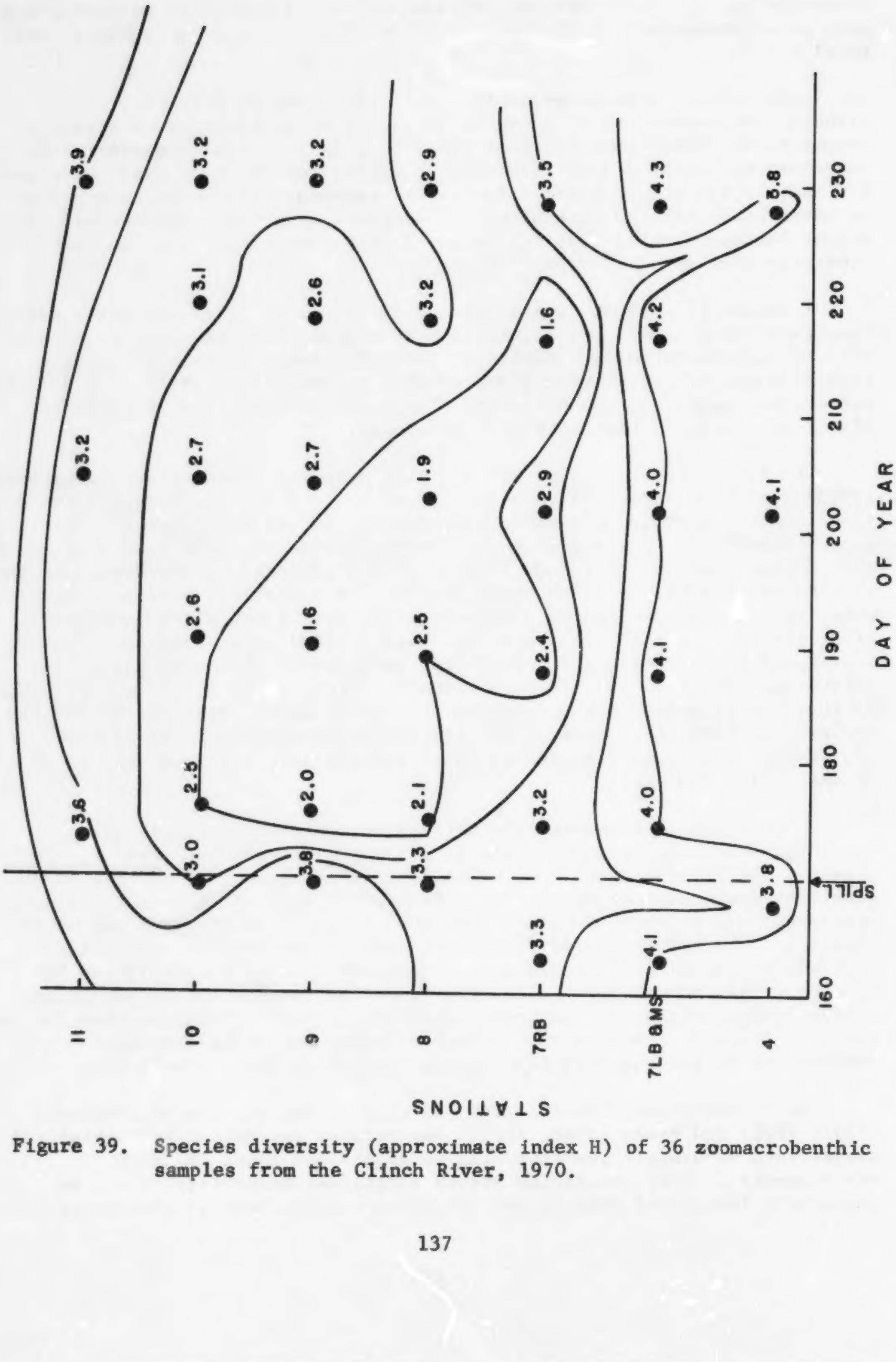


Figure 39. Species diversity (approximate index H) of 36 zoomacrobenthic samples from the Clinch River, 1970.

component for genera (0.52) before the spill was higher than that for families (0.33), but less than that for orders (1.44) (total = 2.29). Immediately after the spill, the difference between familial and generic diversity was even more pronounced: 0.57 (genera), 0.00 (families), 0.79 (orders), and total = 1.36.

This example also suggests that each component of diversity in a taxonomic hierarchy may be a useful indicator of stressed vs. unstressed communities. Note, however, that the diversity for each category in the taxonomic hierarchy did not decrease in proportion to the overall decrease in H. And, if the Clinch River data set is representative, one can conclude that in unstressed, healthy ecosystems the combined components of diversity for higher taxonomic categories (order and family) contribute more to measures of diversity than genera and species.

A second test of the usefulness of hierarchical diversity deals with functional morphology. If all species in a genus have the same functional role in the ecosystem, how important is their diversity to the overall stability of the system when compared with the diversity of other taxa? To answer this question, two hierarchical classifications that bypassed the classical taxonomic hierarchy were developed.

The first classification was modified from the system proposed by Cummins (1973) (refer to Tables 11 through 13, section 4.1.3). It stresses the functional group, feeding mechanism, dependence, and the principal food habit of each species. In adapting it to our needs, numerical codes were assigned to each designation in a category (Table 12) and a hierarchy was developed that is admittedly artifical (Tables 11 and 13). Hierarchical diversity was computed for five samples or subsamples from the Clinch River immediately after the spill (Table 49). Station 7 was divided into substation 7LB&MS, unaffected by the spill and day-to-day operation of the power plant, and substation 7RB, strongly affected by both. Stations 8, 9, and 10 were located in the zone of mixing downstream from the spill site. Stations 7LB&MS and 10 had similar numbers of taxa, but quite different components of diversity at all levels. Other than number of taxa, samples from stations 7RB, 8, 9, and 10 were quite similar.

In addition to summarizing diversity for various levels in each hierarchical classification, Table 49 gives the standard deviations. In this example, the standard deviation was a convenient measure of scatter of the diversity components at each level. For the TF classification, the standard deviation for the control station was much larger than that for any other stations. The high standard deviation found at the control substation resulted primarily from the large component of diversity contributed by species within food habit. Also, the component contributed by dependence within feeding mechanism was appreciably larger for the control than for the affected stations. Substation 7RB had a higher diversity component contributed by feeding mechanism within functional group (Table 50).

The second classification (HBR) resulted from the work of Steinmann (1907, 1908) and Hynes (1960, 1970), who related anatomical and behavioral adaptations of benthic invertebrates to conditions found in lotic environments. Three categories stress functional morphology, i.e., an organism's functional role in the ecosystem and its habitat preference as reflected by its morphology: head position, body shape, and type and shape of respiratory organs (refer to Table 13).

The relative contribution of each component of diversity did not vary appreciably from sample to sample as one proceeded from one category to the next (Table 51). More importantly, each component within the hierarchy contributed to the overall diversity, suggesting that the HBR (head-body-respiratory) classification scheme warrants further investigation.

In the hierarchical diversity study, different classification schemes were used. Hynes (1970) stated, "It should be borne in mind that these are, to a considerable extent, artificial, and that any one species or higher taxon may display several types of adaptation. Alas, some of the phenomena to be discussed are not strictly adaptation; they are changes brought about by life in this peculiar environment." Although it is still too early to discern whether some of the observed trends are real, it appears that classifications emphasizing functional relationships merit further consideration and that hierarchical diversity is a useful heuristic technique.

TABLE 46. SPECIES DIVERSITY (BRILLOUIN'S H) OF 36 ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970

Date	4	7LB&MS	7RB	8	9	10	11
Group I--H values >2.1							
Before pH stress	X	X		X	X	X	
Immediately after pH stress		X					X
2 weeks after		X					
4 weeks after	X	X					X
6 weeks after		X					
8 weeks after	X	X			X	X	X
Group II--H values >1.7 to <2.1							
Before pH stress			X				
Immediately after pH stress			X				
2 weeks after						X	
4 weeks after					X	X	
6 weeks after				X	X	X	
8 weeks after			X	X			
Group III--H values >0.9 to <1.7							
Before pH stress							
Immediately after pH stress				X	X	X	
2 weeks after			X	X	X		
4 weeks after			X	X			
6 weeks after			X				
8 weeks after							

TABLE 47. SPECIES DIVERSITY (APPROXIMATE INDEX H") OF 36 ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970

Date	4	7LB&MS	7RB	8	9	10	11
Group I--$\bar{d}$ values >3.0							
Before pH stress	X	X	X	X	X	X	
Immediately after pH stress		X	X				X
2 weeks after		X					
4 weeks after	X	X					X
6 weeks after		X		X		X	
8 weeks after	X	X	X	X	X	X	X
Group II--$\bar{d}$ values >1 to <3.0							
Before pH stress							
Immediately after pH stress				X	X	X	
2 weeks after			X	X	X	X	
4 weeks after			X	X	X	X	
6 weeks after			X		X		
8 weeks after							

TABLE 48. HIERARCHICAL TAXONOMIC DIVERSITY (BRILLOUIN'S H) OF ZOOMACROBENTHIC SAMPLES FROM THE CLINCH RIVER, 1970

Sample	Period	Order	Component family	Genus	H
4 RB	Before pH stress	1.19	0.97	0.21	2.37
4 LB	Before pH stress	1.55	0.56	0.39	2.50
4 MS	Before pH stress	1.32	0.42	0.32	2.06
4 RB	4 weeks after	1.32	0.75	0.55	2.62
4 LB	4 weeks after	1.28	0.59	0.32	2.19
4 MS	4 weeks after	1.42	0.54	0.44	2.40
4 RB	8 weeks after	1.30	0.63	0.24	2.17
4 LB	8 weeks after	1.20	0.64	0.33	2.17
4 MS	8 weeks after	1.37	0.39	0.35	2.11
7 RB	Before pH stress	1.41	0.47	0.14	2.02
7 LB	Before pH stress	1.37	0.92	0.41	2.70
7 MS	Before pH stress	1.09	1.09	0.25	2.43
7 RB	Immediately after	1.21	0.08	0.50[a]	1.79
7 LB	Immediately after	1.20	0.96	0.29	2.45
7 MS	Immediately after	1.27	0.40	0.52[a]	2.19
7 RB	2 weeks after	0.63	0.06	0.53[a]	1.22
7 LB	2 weeks after	1.35	0.76	0.33	2.44
7 MS	2 weeks after	1.37	0.41	0.45	2.23
7 RB	4 weeks after	0.82	0.10	0.40[a]	1.32
7 LB	4 weeks after	1.21	0.74	0.31	2.26
7 MS	4 weeks after	1.34	0.43	0.34	2.11
7 RB	6 weeks after	0.60	0.12	0.17[a]	0.89
7 LB	6 weeks after	0.83	0.99	0.26	2.08
7 MS	6 weeks after	1.31	0.61	0.19	2.11
7 RB	8 weeks after	1.20	0.18	0.34[a]	1.72
7 LB	8 weeks after	1.26	0.87	0.24	2.37
7 MS	8 weeks after	1.35	0.72	0.22	2.29
8 RB	Before pH stress	1.37	0.29	0.53[a]	2.19
8 LB	Before pH stress	1.44	0.33	0.52[a]	2.29
8 MS	Before pH stress	1.39	0.42	0.54[a]	2.35

TABLE 48 (continued)

Sample	Period	Order	Component family	Genus	H
8 RB	Immediately after	0.87	0.22	0.44	1.33
8 LB	Immediately after	0.79	0.00	0.57[a]	1.36
8 MS	Immediately after	0.84	0.07	0.52[a]	1.43
8 RB	2 weeks after	0.61	0.05	0.95[a]	1.61
8 LB	2 weeks after	0.62	0.04	0.94[a]	1.60
8 MS	2 weeks after	0.49	0.01	0.89[a]	1.39
8 RB	4 weeks after	1.32	0.75	0.55	2.62
8 LB	4 weeks after	1.28	0.59	0.32	2.19
8 MS	4 weeks after	1.42	0.54	0.44	2.40
8 RB	6 weeks after	0.56	0.05	0.62[a]	1.23
8 LB	6 weeks after	1.01	0.12	0.65[a]	1.78
8 MS	6 weeks after	0.86	0.07	0.64[a]	1.57
8 RB	8 weeks after	1.19	0.25	0.42[a]	1.86
8 LB	8 weeks after	1.33	0.33	0.17	1.83
8 MS	8 weeks after	1.15	0.16	0.40[a]	1.71
9 RB	Before pH stress	1.41	0.33	0.64[a]	2.38
9 LB	Before pH stress	1.40	0.69	0.38	2.47
9 MS	Before pH stress	1.21	0.19	0.44[a]	1.84
9 RB	Immediately after	0.67	0.03	0.65[a]	1.35
9 LB	Immediately after	0.43	0.13	0.83[a]	1.39
9 MS	Immediately after	0.85	0.06	0.71[a]	1.62
9 RB	2 weeks after	0.28	0.03	0.75[a]	1.06
9 LB	2 weeks after	0.97	0.09	0.81[a]	1.87
9 MS	2 weeks after	0.71	0.01	0.80[a]	1.52
9 RB	4 weeks after	0.97	0.14	0.69[a]	1.80
9 LB	4 weeks after	0.90	0.14	0.53[a]	1.57
9 MS	4 weeks after	1.07	0.17	0.60[a]	1.84
9 RB	6 weeks after	1.39	0.12	0.39[a]	1.90
9 LB	6 weeks after	1.29	0.20	0.44[a]	1.93
9 MS	6 weeks after	1.31	0.05	0.16[a]	1.52
9 RB	8 weeks after	1.38	0.17	0.50[a]	2.05
9 LB	8 weeks after	1.58	0.31	0.14	2.03
9 MS	8 weeks after	1.56	0.24	0.10	1.90

TABLE 48 (continued)

Sample	Period	Order	Component family	Genus	H
10 RB	Before pH stress	1.20	0.29	0.38[a]	1.87
10 LB	Before pH stress	1.46	0.62	0.31	2.39
10 MS	Before pH stress	1.32	0.36	0.35	2.03
10 RB	Immediately after	1.11	0.04	0.44[a]	1.59
10 LB	Immediately after	1.25	0.06	0.48[a]	1.79
10 MS	Immediately after	0.86	0.03	0.37[a]	1.26
10 RB	2 weeks after	1.19	0.18	0.40[a]	1.77
10 LB	2 weeks after	1.20	0.20	0.33[a]	1.73
10 MS	2 weeks after	1.33	0.39	0.33	2.05
10 RB	4 weeks after	1.28	0.19	0.35[a]	1.82
10 LB	4 weeks after	1.15	0.20	0.33[a]	1.68
10 MS	4 weeks after	1.14	0.11	0.34[a]	1.59
10 RB	6 weeks after	1.36	0.15	0.50[a]	2.01
10 LB	6 weeks after	1.24	0.19	0.34[a]	1.77
10 MS	6 weeks after	1.23	0.22	0.27[a]	1.72
10 RB	8 weeks after	1.42	0.24	0.46[a]	2.12
10 LB	8 weeks after	1.63	0.35	0.29	2.27
10 MS	8 weeks after	1.36	0.39	0.16	1.91
11 RB	Immediately after	1.34	0.71	0.24	2.29
11 LB	Immediately after	1.12	0.78	0.26	2.16
11 MS	Immediately after	1.24	0.51	0.35	2.10
11 RB	4 weeks after	1.17	0.49	0.36	2.02
11 LB	4 weeks after	1.22	0.46	0.39	2.07
11 MS	4 weeks after	1.30	0.57	0.37	2.24
11 RB	8 weeks after	1.37	0.85	0.17	2.38
11 LB	8 weeks after	1.42	0.63	0.18	2.23
11 MS	8 weeks after	1.33	0.67	0.26	2.26

[a] The generic component of diversity contributed more to H than the familial component of diversity.

TABLE 49. COMPONENT OF SPECIES DIVERSITY (H) AT EACH LEVEL IN THE TROPHIC-FUNCTIONAL HIERARCHY FOR FIVE SAMPLES OR SUBSAMPLES COLLECTED IMMEDIATELY AFTER THE ACID SPILL ON THE CLINCH RIVER, 1970

		Additive component						
Station	Taxa	Functional group	Feeding mechanism within functional group	Dependence within feeding mechanism	Food habit within dependence	Species within food habit	H	Standard deviation
7LB&MS	44	0.90	0.36	0.29	0.03	1.16	2.74	0.47
7RB	12	0.70	0.32	0.12	0.00	0.58	1.72	0.30
8	28	0.82	0.09	0.04	0.00	0.65	1.60	0.38
9	24	0.78	0.12	0.02	0.03	0.74	1.69	0.39
10	39	0.89	0.09	0.03	0.01	0.54	1.65	0.38

TABLE 50. PERCENT OF SPECIES DIVERSITY (H) CONTRIBUTED AT EACH LEVEL IN THE TROPHIC-FUNCTIONAL HIERARCHY FOR FIVE SAMPLES OR SUBSAMPLES COLLECTED IMMEDIATELY AFTER THE ACID SPILL ON THE CLINCH RIVER, 1970

		Additive component				
Station	Taxa	Functional group	Feeding mechanism within functional group	Dependence within feeding mechanism	Food habit within dependence	Species within food habit
7LB&MS	44	32.7	13.0	10.6	1.0	42.6
7RB	12	40.9	18.2	6.9	0.0	34.0
8	28	51.2	5.9	2.3	0.0	40.6
9	24	46.0	7.0	1.4	1.9	43.7
10	39	53.9	5.6	7.6	0.1	32.8

TABLE 51. COMPONENT OF SPECIES DIVERSITY (H) AT EACH LEVEL IN THE HEAD-BODY-RESPIRATORY FUNCTIONAL MORPHOLOGY HIERARCHY FOR FIVE SAMPLES OR SUBSAMPLES COLLECTED IMMEDIATELY AFTER THE ACID SPILL ON THE CLINCH RIVER, 1970

	Additive component					
Station	Head position	Body shape within head position	Respiratory organ within body shape	Species within respiratory organ	H	Standard deviation
7LB&MS	0.67	1.29	0.84	0.61	2.74	0.87
7RB	0.63	0.29	0.36	0.44	1.72	0.47
8	0.70	0.19	0.09	0.62	0.60	1.39
9	0.69	0.19	0.08	0.73	1.69	0.44
10	0.64	0.43	0.09	0.46	1.62	0.40

SECTION 8

SUMMARY AND DISCUSSION

8.1 INTRODUCTION

Different methods of analysis were used with varying results. The purpose of this section is to evaluate the results obtained, to assess the different levels of interpretability, and to make general recommendations about the application of these methods and the results one can expect from them.

8.2 NATURE OF THE ECOSYSTEMS FROM WHICH DATA BASES WERE SELECTED

A point that must be stressed is that the nature of the ecosystem and the nature of the data collected from the ecosystem determine both the methods to be used in analyzing the data and the ease with which the results can be interpreted. For example the Clinch River is a shallow, rapidly flowing stream with a normally diverse benthic community that was adversely impacted by an acute pH stress. Numerous large samples were collected at closely-timed intervals both before and after the acute pH stress. Like many streams, it could be regarded as a linear system, in which contemporaneous samples from adjacent stations are expected to be similar. The Cumberland River, on the other hand, was studied in an impounded area with deep, slowly moving water that received chronic thermal stress. Although both high- and low-flow conditions were studied, the velocity was never as great as the Clinch River. Moreover, although temperatures in the range of the Cumberland River are common in nature and adapted to by numerous warm-water organisms, the low pH stress received by the Clinch River was a shock that few organisms were able to tolerate. Finally, the Cumberland River was not a linear system, and the thermal plume (potential stress) moved upstream during times of low flow.

The depth of water in the Clinch River was much less than the depth in the Cumberland River, where benthic stations were located at depths varying from 2.5 to 9 meters. Samples from the Clinch River contained more individuals and greater diversities than samples from the Cumberland River. This difference stems in part from more extensive sampling in the Clinch and in part from real differences between the two ecosystems. The relative uniformity of the substratum sampled in the Cumberland River provided fewer data from fewer habitats and niches. Depth reduced the impact of the stress in the Cumberland, whereas all macroinvertebrates in the Clinch River downstream from the site of the acid spill were exposed to the low pH shock. The net effect of the differences between the ecosystems and the kind and degree of stress resulted in the Clinch River ecosystem being more understandable and definable than the Cumberland River system.

While major environmental changes are evident with coarse methods of analysis, subtle impacts may go undetected even with elaborately designed studies. This is especially true when unexpected factors, such as upstream flow of the plume, inhibit the effectiveness of the analytical protocol established for the study. In our examples, the Clinch River was studied in greater detail than necessary to determine initial impact and recovery. The Cumberland River was probably not sampled adequately in view of the problem caused by upstream flow. The important point is that it is virtually impossible to detect most subtleties or the lack of them before the fact. Thus, the adequacy of a sampling program may be in doubt until after some of the data has been analyzed. Preliminary studies are too rarely conducted. As a result judgmental errors may be made that limit the value of the study or seriously overextend resources.

8.3 METHODS

8.3.1 Relationships Between Methods

Cluster analysis and ordination of data are intended to serve the same purpose when used in the Q-mode, i.e., to express similarities between samples from different stations. Similarities between samples are based on their faunal content with species given equal weight a priori. Because cluster analysis will force samples into clusters whether or not clusters exist in nature, it may be unsuited for analysis of strictly linear (e.g., riverine) systems. When numerous samples are collected through time, however, as was done in study of the Clinch River, the strict linearity of the system is overcome due to the imposition of temporal trends. In such cases, cluster analysis may be a useful analytical tool. For river-reservoir systems, such as the Cumberland River, where overbank areas and channel areas may have high intracorrelations regardless of their position along the river, cluster analysis may be entirely appropriate.

Throughout this report we have stressed cluster analysis and have deemphasized ordination because the application and interpretation of cluster analysis to problems of applied aquatic ecology are much better established. Specifically, procedures have not been developed and tested for limiting the size of data sets to be studied by ordination and for excluding rare species that may be present in some samples and absent from others due to chance alone. In spite of being somewhat more difficult to interpret, ordination methods such as nonmetric multidimensional scaling can show everything cluster analysis shows without some of the inherent disadvantages. Ordination methods merit further investigation by ecologists.

Analysis of species diversity serves a different purpose from cluster analysis and ordination. Indices of species diversity do not consider what species are present. They are based on the number of species present and the evenness of the distribution of individuals among species. Thus, species diversity is a measure of community structure that is independent of the particular community being studied. It is tempting to compute indices of diversity from numerous communities and to compare the state of health of the ecosystems by comparing indices. Three aspects of communities suggest that this approach must be used with caution if at all. First, not all natural communities have high diversity. For example, undisturbed communities of macroinvertebrates from rapidly flowing streams are more diverse than those from slower moving streams and reservoirs. The difference in diversity is not a difference in health of the ecosystems, although it may bear on their resiliency. Second, undisturbed upstream communities may be as diverse as undisturbed downstream ones, but the faunas may be quite different. Third, a diverse community that is tolerant to heated water or other pollutional stress may become established . Such a community may be described as perched if it is faunally different from upstream communities that would ordinarily contribute to repopulation of the area. An incremental increase in pollutional stress or the temporary removal of stress may decimate the perched community and leave it degraded since no source of stress-tolerant species exists upstream. None of these three points can be considered adequately by the rote comparison of species diversity alone.

8.3.2 Data

The kind and nature of data dictate the type of procedures that should be used for a study, the conclusions that can be drawn, and the confidence one has in those conclusions. The most important aspect of planning any ecological study is to make sure the data collected are adequate for answering the question at hand. Sampling must have a well defined purpose and the more clearly that purpose is stated the greater the likelihood that data will be adequate for the intended purpose. In Section 7 we dealt briefly with problems that may arise from collecting too small a sample. Other aspects of sampling design should also be considered to develop sampling protocols that produce data of the desired quality and quantity.

8.3.2.1 Presence-Absence Data--

The use of presence-absence data has one serious drawback: all information contributed by the differential abundances of organisms is lost. For some purposes, however, the benefits to be gained may outweigh the disadvantages. First, presence-absence data can be obtained much more quickly, and large samples can be processed at less expense. Second, depending on the purpose of the study, presence-absence data may give all the information needed. It was also determined that quantitative data collected during a cursory survey or poorly documented field work should be used only as presence-absence data. Note also that a wide selection of coefficients of similarity are available for comparing stations and assembling groups of species by cluster analyses.

Of the numerous coefficients available for cluster analyses of presence-absence data, two were tested: Jaccard's coefficient, which omits negative matches, and the simple matching coefficient, which regards negative matches (the absence of a species from two stations) as contributing to similarity. Both coefficients range from 0 to 1. In practice, however, Jaccard's coefficients are lower than simple matching coefficients because data from aquatic environmental surveys are often characterized by a large number of species with few individuals that differ from station to station. As a result mutual absences (negative matches) may indicate more similarity between samples than warranted. To minimize this effort we recommend Jaccard's coefficient except for those rare cases where mutual absences clearly result from the same cause. If, however, a species is absent from one station because the temperature is too high and absent from another because the substrate is unsuitable, the simple matching coefficient may cause an investigator to draw an erroneous conclusion. If Jaccard's coefficients are uniformly low, one should be wary of randomness or absence of species by chance alone due to sampling error.

8.3.2.2 Quantitative Data--

Quantitative data usually consist of the numbers of individuals per species in a sample. Sometimes not all groups of organisms are distinguished at the species level, so the counts may be expressed as numbers of individuals at higher taxonomic levels. Specifically, oligochaetes and chironomids are difficult, costly, and time consuming to identify as to species and are often lumped. Quantitative data may also be reported as ranked abundances or as proportions of each species or higher taxon in a sample.

Individual species counts are sometimes transformed before analyses are performed. Transformations are usually intended to reduce the quantitative impact of highly abundant species. Ranking data is a kind of transformation that is roughly equivalent to using a logarithmic transformation. Either procedure is drastic. The logarithmic transformation, for example, changes abundances of 1, 10, 100, and 1000 to 0, 1, 2, and 3. The square-root transformation is less drastic ($\sqrt{y + 0.5}$), changing abundances of 1, 10, 100, and 1000, respectively, to 1.22, 3.24, 10.02, and 31.63. This transformation worked well with all three data sets studied. Finally, counts may be standarized, whereby a data matrix of $\underline{t}$ samples (columns) and $\underline{n}$ species (rows) is operated on, such that each element in a row has the row mean subtracted from it and is divided by the row standard deviation to give a new row mean of 0 and a new row standard deviation of 1. This procedure is effective in removing inordinate effects of species that are very abundant throughout a study. Use of data with species abundances expressed as proportions was not examined in this report. Because of problems that may arise with proportions, the arcsine transformation (arcsine $\sqrt{p}$) is recommended for proportional data. We suggest that use of proportional data be tested in the future.

Two coefficients for use in cluster analysis of quantitative data were tested: the Pearson product-moment correlation coefficient and Sokal's taxonomic distance coefficient. In general, the correlation coefficient shows high similarities between samples with species present in the same relative proportions, whereas the distance coefficient shows high similarity (low distance) between samples with species present in the same absolute abundances. Either coefficient produced useful results with data transformed by the square-root transformation and unstandardized. Of course, since the two coefficients measure different things, the results were different; and, in general, cluster analysis of correlation coefficients proved to be more readily interpretable.

8.3.3 Cluster Analysis

Cluster analysis was computed in two modes: Q-mode, which shows similarities among samples on the basis of their contained species; and R-mode, which shows similarities among species on the basis of their distributions and abundances in samples. Q-mode cluster analysis has been used much more than R-mode in applied aquatic ecology. Results of Q-mode analysis are applicable only to the stream or reservoir from which data have been collected and clustered. R-mode analysis, on the other hand, has promise as a method of comparing faunal associations from stream to stream and from basin to basin.

Fortunately, the distortion introduced during clustering can be assessed by the coefficient of cophenetic correlation (r_{cc}). Cluster analysis not accompanied by such a coefficient is suspect, and we have used an $r_{cc} \geq 0.8$ as the criterion for consideration in data interpretation. Once it has been determined that distortion is acceptably low, the overall shape and the levels of clustering of a dendrogram provide useful information about the data being clustered and the reliability of interpretations. Ideally a dendrogram should have several good, tight clusters that are distinct from other clusters or samples in the dendrogram (see, e.g., Figures 11 and 15). Such dendrograms have sufficient structure that they may be easily interpreted, depending, of course, on the samples that comprise the clusters. An example of a dendrogram in which the similarities are all rather low and of about the same value is Figure 12. One interpretation is that similarities among samples are actually low and more or less equal throughout the study. A common cause of such clusters, however, is randomness. In Figure 12, for example, all species were included in the data matrix, and no attempt was made to remove rare species that were present in some samples and, perhaps, absent from others by chance alone due to sampling error. When such data are used with Jaccard's coefficient, which always gives low similarities because it ignores negative matches, the resulting similarities can have a large, unknowable component of randomness. Another cause of the shape of such dendrograms can be lack of natural clusters. Recall that cluster analysis forces samples into clusters whether or not such clusters exist in nature. If the system sampled is strictly linear with sequential similarities such as might be expected in a stream, the resultant dendrogram can lack structure. (To use cluster analysis with such data is analogous to trying to logically subdivide a piece of string. In such cases ordination is called for as an alternative to cluster analysis.) Other dendrograms have a stair-stepped arrangement and shape (see e.g., Figure 31). Such shape usually implies the lack of real clusters in nature and again calls for ordination.

In our examples, we have usually chosen a single level of similarity for comparison of all clusters in a dendrogram. In practice, there is no compelling reason why more than one level cannot be chosen, with a different level of similarity used in different parts of the dendrogram. Samples from the high-gradient, upstream portion of a stream, for example, may generally be more similar to each other than samples from the low-gradient, downstream portion. Similarly, samples collected in winter may generally be more similar to each other than samples collected in summer. As with analysis of species diversity, a thorough understanding of the ecosystem and a grasp of the questions to which answers are being sought is required, not mere cookbook applications of complex quantitative analytical techniques.

In some dendrograms, especially those with low overall similarities, a number of samples may be left unclustered at any reasonable level of similarity. In Figure 19, for example, six of the lower eight samples in the dendrogram are unclustered. Lack of clustering can result from real dissimilarity or can be an artifact of clustering. One should refer to the original similarity matrix to check the values and, if necessary, go back to the data matrix to see which species contributed to the anomalous lack of clustering of some samples.

Cluster analysis is best for expressing similarities at the tips of dendrograms (smaller pair groups, higher similarity levels) and is less reliable for interpretation of intercluster similarities (lower similarity levels). Reliability is lost because the clustering algorithm averages similarities at each step in the clustering procedure. If one is interested in intergroup similarities, he is well advised to use principal component ordination, which was not considered in this report.

8.3.4 Ordination

As was pointed out earlier, ordination has been little used in applied aquatic ecology, and our evaluation of it here is mainly in terms of how well it agrees with the results of cluster analysis. In general, however, ordination is a more widely applicable technique than cluster analysis because no a priori assumptions need be made about clusters in the data. Such ordination techniques as principal component analysis are well suited for showing intercluster similarities rather than fine details of intracluster similarities. The use of nonmetric multidimensional scaling tested was a compromise between principal component analysis and cluster analysis. We have found that when no structure exists in a data matrix due to small sample sizes or randomness, nonmetric multidimensional scaling is quick to show it by producing totally meaningless and uninterpretable ordination. We recommend that more work be done with ordination of other data sets to test the applicability of nonmetric multidimensional scaling to data from aquatic surveys.

8.3.5 Species Diversity and Hierarchical Diversity

Because species diversity is meant to provide a different measure of community structure than cluster analysis or ordination, we suggest that it be used in conjunction with the other methods. Indices of species diversity have been widely misused, and a vast literature discusses the disadvantages of using them. The principal difficulty stems from the nonuniqueness of a given value of species diversity. Thus, a moderate value of species diversity may result from a wide range of community structures. Communities with many species with few individuals each, a moderate number of species with a moderate number of individuals each, and few species with many individuals each could all give the same value of species diversity, depending on the sampling and evenness of distribution of individuals among species. With such latitude, the method must of course be used with care. In particular, the search for absolute, global values of species diversity to indicate healthy or damaged ecosystems is wrong in principal. Nevertheless, species diversity of samples from within a stream, reservoir, or drainage basin can be compared, especially if used with care and in conjunction with cluster analysis or ordination.

Hierarchical diversity has been little used in applied aquatic ecology except by authors of this report and their coauthors. We recommend that it be further explored because of its value in showing ways in which the high cost of aquatic environmental surveys can be reduced.

REFERENCES

Barr, A. J., J. H. Goodnight, J. P. Sall, and J. T. Helwig. 1976. A User's Guide to SAS·76. SAS Institute, Inc., Raleigh, North Carolina. Sparks Press pp 329.

Basharin, G. P. 1959. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory Probab*. *Its Appl*. 4:333-36.

Brillouin, L. 1962. *Science and Information Theory*. 2nd ed. New York: Academic Press. 347 pp.

Buchanan, R. J. and B. Lighthart. 1973. Indicator Phytoplankton Communities: A Cluster Analysis Approach. *B. C. Prov. Mus*. *Nat. Hist. Anthropol. Rep*. 6:1-10.

Cairns, John, Jr., D. W. Albaugh, F. Busey, and M. D. Chanany. 1968. The Sequential Comparison Index--A Simplified Method for Nonbiologists to Estimate Relative Differences in Biological Diversity in Stream Pollution Studies. *J. Water Pollut. Control Fed*. 40: 1607-1613.

Cairns, J., Jr., and R. L. Kaesler. 1969. Cluster Analysis of Potomac River Survey Stations Based on Protozoan Presence-Absence Data. *Hydrobiol*. 34:414-32.

Cairns, J., Jr., and R. L. Kaesler. 1971. Cluster Analysis of Fish in a Portion of the Upper Potomac River. *Trans. Am. Fish Soc*. 100:750-56.

Cairns, J., Jr., R. L. Kaesler, and R. Patrick. 1970. Occurrence and Distribution of Diatoms and Other Algae in the Upper Potomac River. *Nat. Acad. Nat. Sci. Philadelphia* 436:1-12.

Cairns, J., Jr., G. R. Lanza, and B. C. Parker. 1972. Pollution Related Structural and Functional Changes in Aquatic Communities with Emphasis on Freshwater Algae and Protozoa. *Proc. Acad. Nat*. *Sci. Philadelphia* 124(5):79-127.

Cheetham, A. H. and J. E. Hazel. 1969. Binary (Presence-Absence) Similarity Coefficients. *J. Paleontol*. 43(5):1130-36.

Crossman, J. S., J. Cairns, Jr., and R. L. Kaesler. 1973. Aquatic Invertebrate Recovery in the Clinch River Following Hazardous Spills and Floods. *Va. Polytech. Inst. State Univ. Water Resour. Res. Cent. Bull*. 63:1-56.

Crossman, J. S., R. L. Kaesler, and J. Cairns, Jr. 1974. The Use of Cluster Analysis in the Assessment of Spills of Hazardous Materials. Am. Midl. Nat. 92(1):94-114.

Cummins, K. W. 1973. Trophic Relations of Aquatic Insects. Annu. Rev. Entomol. 18:183-206.

Dennison, J. M. and W. W. Hay. 1967. Estimating the Needed Sampling Area for Subaquatic Ecologic Studies. J. Paleont. 41:706-708.

Environmental Protection Agency. 1976. Federal Interagency Energy/Environment Research and Development Program, Status Report II. Office of Energy, Minerals, and Industry, Office of Research and Development, Washington, DC, 1976.

Farris, J. S. 1969. On the Cophenetic Correlation Coefficient. Syst. Zool. 18:279-85.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J. Anim. Ecol. 12:42-58.

Forbes, S. A. 1907. On the Local Distribution of Certain Illinois Fishes: An Essay in Statistical Ecology. Bull. Ill. State Lab. Nat. Hist. 7:273-303.

Green, P. E. and F. J. Carmone. 1970. Multidimensional Scaling and Related Techniques in Marketing Analysis. Boston, Mass.: Allyn and Bacon, Inc. 203 pp.

Green, R. H., Jr. 1979. Sampling Design and Statistical Methods for Environmental Biologists. Wiley-Interscience, New York, 257 pp.

Hamilton, M. A. 1975. Indexes of Diversity and Redundancy. J. Water Pollut. Control Fed. 47:630-32.

Hedgpeth, Joel. 1973. Temperature Relationships of Near Shore Oceanic and Estuarine Communities. In Effects and Methods of Control of Thermal Discharges, pp. 1271-1431. Serial No. 93-14, Pt. 3, Washington, DC: U.S. Govt. Printing Office.

Hoaglin, David C. and Roy E. Welch. 1975. MIT-SNAP. An Interactive Data Analysis Systems. MIT: 60 pages.

Hurlbert, S. H. 1971. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology 52:577-86.

Hutchinson, G. E. 1957. Concluding Remarks, Cold Spring Harbor Symposium. Quant. Biol. 22:415-27.

Hynes, H.B.N. 1960. Biology of Polluted Waters. Liverpool: Liverpool Univ. Press. 202 pp.

Hynes, H.B.N. 1970. The Ecology of Running Waters. Toronto: Univ. of Toronto Press. 555 pp.

Jaccard, P. 1908. Nouvelles Recherches sur la Distribution Florale. So. Vanoise Sci. Natur. Bull. 44:233-70.

Kaesler, R. L. 1970. The Cophenetic Correlation Coefficient in Paleo-Ecology. Bull. Geol. Soc. Am. 81:1261-66.

Kaesler, R. L. and J. Cairns, Jr. 1972. Cluster Analysis of Data from Limnological Surveys of the Upper Potomac River. Am. Midl. Nat. 88:56-67.

Kaesler, R. L., J. Cairns, Jr., and J. M. Bates. 1971. Cluster Analysis of Noninsect Macroinvertebrates of the Upper Potomac River. Hydrobiol. 37(2):173-81.

Kaesler, R. L., J. Cairns, Jr., and J. S. Crossman. 1974. Redundancy in Data from Stream Surveys. Water Res. 8:637-42.

Kaesler, R. L. and E. E. Herricks. 1977. Analysis of Data from Biological Surveys of Streams: Diversity and Sample Size. Water Res. Bull. 13(1):125-35.

Kaesler, R. L., E. E. Herricks, and J. S. Crossman. 1978. Uses of Indices of Diversity and Hierarchical Diversity in Stream Surveys. In ASTM Symposium on Quantitative and Statistical Analyses of Biological Data for the Assessment of Water and Wastewater Quality, Minneapolis, Minn. (June 20-21, 1977).

Kolkwitz, R. and M. Marrson. 1908. Ecology of Plant Saprobia. Rep. Ger. Bot. Soc. 26a:505-19.

Kolkwitz, R. and M. Marrson. 1909. Ecology of Animal Saprobia. Int. Rev. Hydrobiol. Hydrogeogr. 2:126-52.

Krebs, C. J. 1972. Ecology: The Experimental Analysis of Distribution and Abundance. New York: Harper and Row. 694 pp.

Kruskal, J. B. 1964a. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika. 29:1-27.

Kruskal, J. B. 1964b. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika. 219:115-29.

Margalef, R. 1956. Information y Diversidad Espicfica en las Cominudades de Organismas. Invest. Pesq. 3:99-106.

Parker, B. C. and B. L. Turner. 1961. Operational Niches and Community-Interaction Values as Determined from In Vitro Studies of Some Soil Algae. Evolution. 15(2):228-238.

Patil, G. P. and C. Taillie. 1976. Ecological Diversity: Concepts, Indices, and Applications. Proc. 9th Int. Biom. Conf. 2:383-411.

Patrick, R. 1961. A Study of the Numbers and Kinds of Species Found in Rivers in Eastern United States. *Proc. Acad. Nat. Sci. Philadelphia* 113(10):215-58.

Patrick, R. 1967. Natural and Abnormal Communities of Aquatic Life in Rivers. *Bull. S. C. Acad. Sci*. 29:19-28.

Peet, R. K. 1974. The Measurement of Species Diversity. *Ann. Rev. Eco. Syst*. 5:285-307.

Pennak, R. W. 1971. Toward a Classification of Lotic Habitats. *Hydrobiol*. 30:321-334.

Pielou, E. C. 1966a. The Measurement of Diversity in Different Types of Biological Collection. *J. Theor. Biol*. 13:131-44.

Pielou, E. C. 1967. The Use of Information Theory in the Study of the Diversity of Biological Population. *Proc. 5th Berkeley Symp. Math. Stat. Probab*. 4:163-77.

Pielou, E. C. 1969. *An Introduction to Mathematical Ecology*. New York: John Wiley & Sons. 286 pp.

Pielou, E. C. 1974. *Population and Community Ecology: Principles and Methods*. New York: Gordon and Breach. 424 pp.

Pielou, E. C. 1975. *Ecological Diversity*. New York: Wiley-Interscience. 165 pp.

Pielou, E. C. 1977. *Mathematical Ecology*. Wiley-Interscience, New York, 311 pp.

Roback, S. S., J. Cairns, Jr., and R. L. Kaesler. 1969. Cluster Analysis of Occurrence and Distortion of Insect Species in a Portion of the Potomac River. *Hydrobiol*. 34:484-502.

Rohlf, F. J. 1970. Adaptive Hierarchical Cluster Schemes. *Syst. Zool*. 19:58-82.

Rohlf, F. J. 1972. An Empirical Comparison of Three Ordination Techniques in Numerical Taxonomy. *Syst. Zool*. 21:271-280.

Shannon, C. E. and W. Weaver. 1949. *The Mathematical Theory of Communication*. Urbana: Univ. of Illinois Press.

Shannon, E. E. 1970. *Eutrophication-Trophic State Relationships in North and Central Florida Lakes*. Ph.D. Thesis, Univ. of Florida. 258 pp.

Shelford, V. E. 1915. Principles and Problems of Ecology as Illustrated by Animals. *J. Ecol*. 3:1-23.

Simpson, E. H. 1949. Measurement of Diversity. Nature 163:688.

Simpson, G. G. 1960. Notes on the Measurement of Faunal Resemblance. Am. J. Sci. 258a:300-11.

Sneath, P.H.A. and R. Sokal. 1973. Numerical Taxonomy. San Francisco: W. H. Freeman and Co. 573 pp.

Sokal, R. R. 1961. Distance as a Measure of Taxonomic Similarity. Syst. Zool. 10:70-79.

Sokal, R. R. and F. J. Rohlf. 1962. The Comparison of Dendrograms by Objective Methods. Taxon 11(2):33-40.

Sokal, R. R. and P.H.A. Sneath. 1963. Principles of Numerical Taxonomy. San Francisco: W. H. Freeman and Co. 359 pp.

Soukup, J. F. 1970. Fish Kill # 70-025, Clinch River, Carbo, Russell County. Unpublished.

Steinmann, P. 1907. Die Tierwelt der Gebirgsbache: Eine Faunistisch-Biologische Studie. Ann. Biol. Lacustre 2:30-150.

Steinmann, P. 1908. Die Tierwelt der Gebirgsbache. Arch. Hydrobiol. 3:266-73.

Stephenson, W. 1972. The Use of Computers in Classifying Marine Bottom Communities. Oceanogr. South Pac. 31:463-73.

Stephenson, W. and M.C.L. Dredge. 1976. Numerical Analysis of Fish Catches from Serpentine Creek. Proc. R. Soc. 87:33-43.

Stephenson, W., Y. I. Raphael, and S. D. Cook. 1976. The Macrobenthos of Bramble Bay, Moreton Bay, Queensland. Mem. Od. Mus. 17(3):425-47.

Stephenson, W., W. T. Williams, and S. D. Cook. 1972. Computer Analysis of Petersen's Original Data on Bottom Communities. Ecol. Monogr. 42(4):387-415.

Train, Russel E. 1973. Address to the National Conference on Managing the Environment. U.S. EPA.

Whittaker, R. H. 1975. Communities and Ecosystems. 2nd ed. New York: Macmillan. pp. 124-25.

Whitten, B. A., ed. 1975. River Ecology. Berkeley: Univ. of Calif. Press. 725 pp.

Wilhm, J. L. 1968. Use of Biomass Units in Shannon's Formula. Ecology 49:153-156.

Wilhm, J. L. and T. C. Dorris. 1968. Biological Parameters for Water Quality Criteria. Bioscience 18(6):477-81.

Williams, C. B. 1950. The Application of the Logarithmic Series to the Frequency of Occurrence of Plant Species in Quadrats. J. Ecol. 38:107-138.

Woodwell, G. M. 1970. Effects of Pollution on the Structure and Physiology of Ecosystems. Science 168:429-33.